

機械学習を自然現象の理解・発見に使いたい人に 知っておいてほしいこと

2021年12月10日

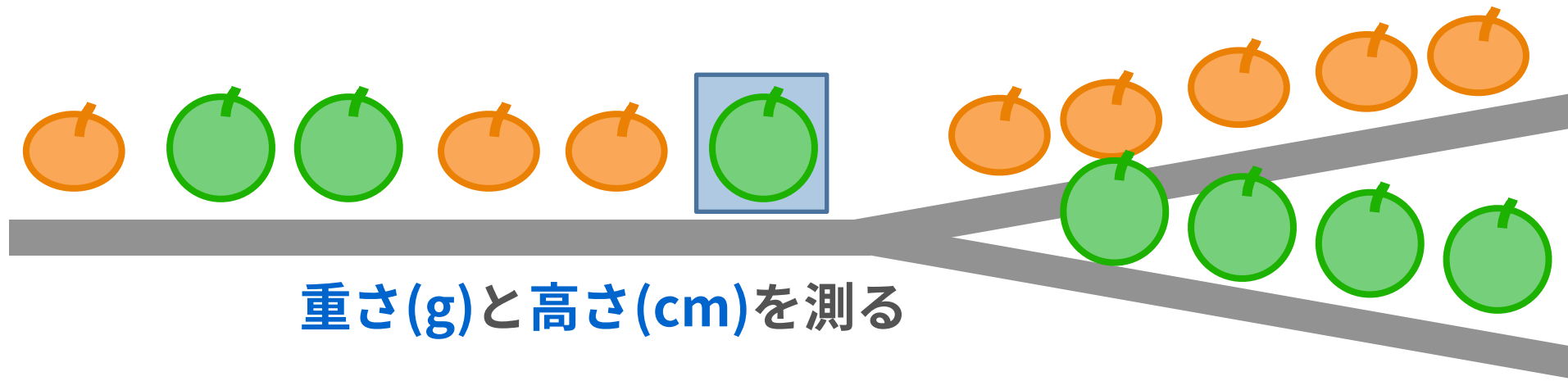
瀧川 一学

ichigaku.takigawa@riken.jp

理化学研究所 革新知能統合研究センター
北海道大学 化学反応創成研究拠点 (ICReDD)

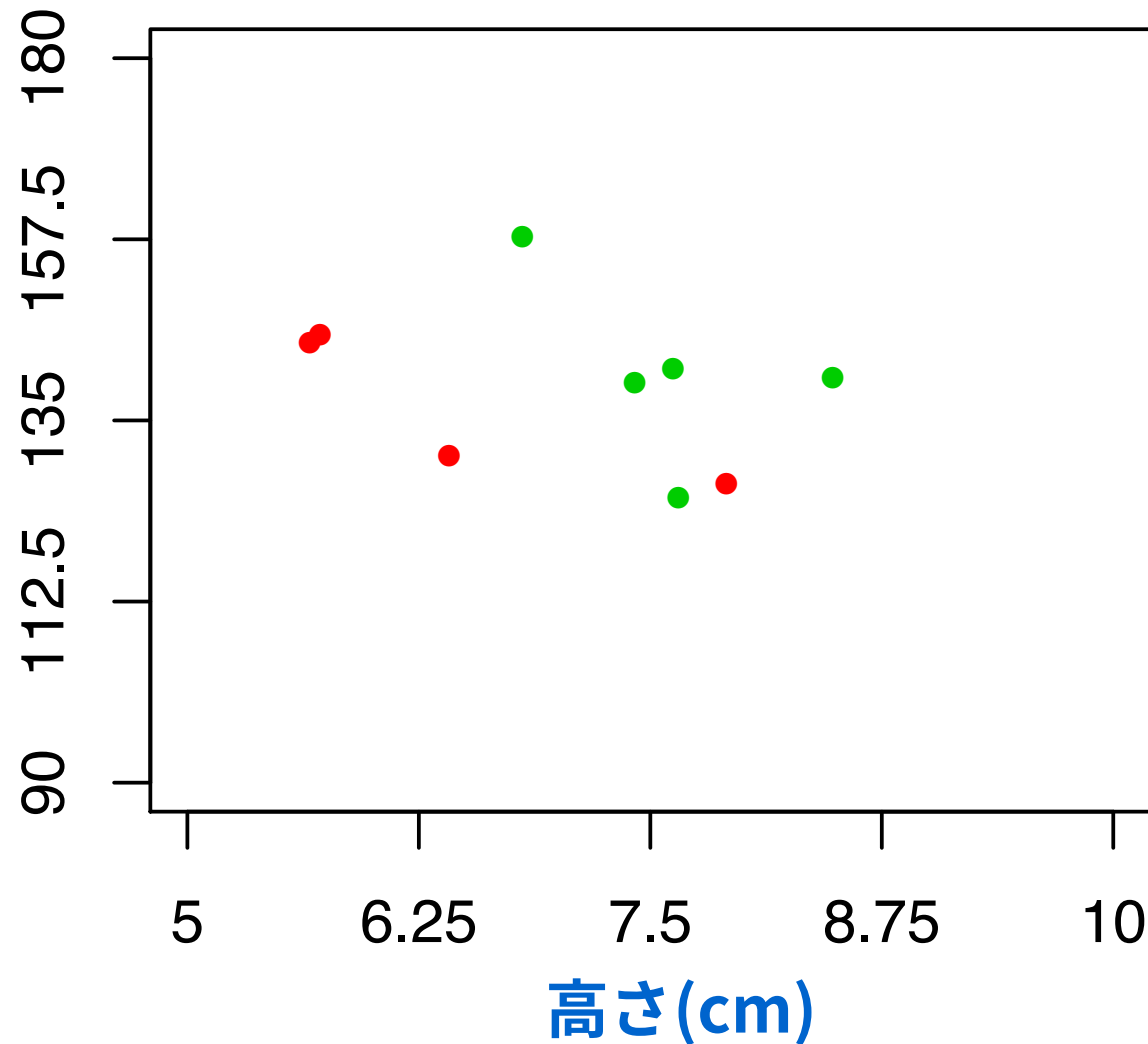


機械学習は「データを予測に変える」技術

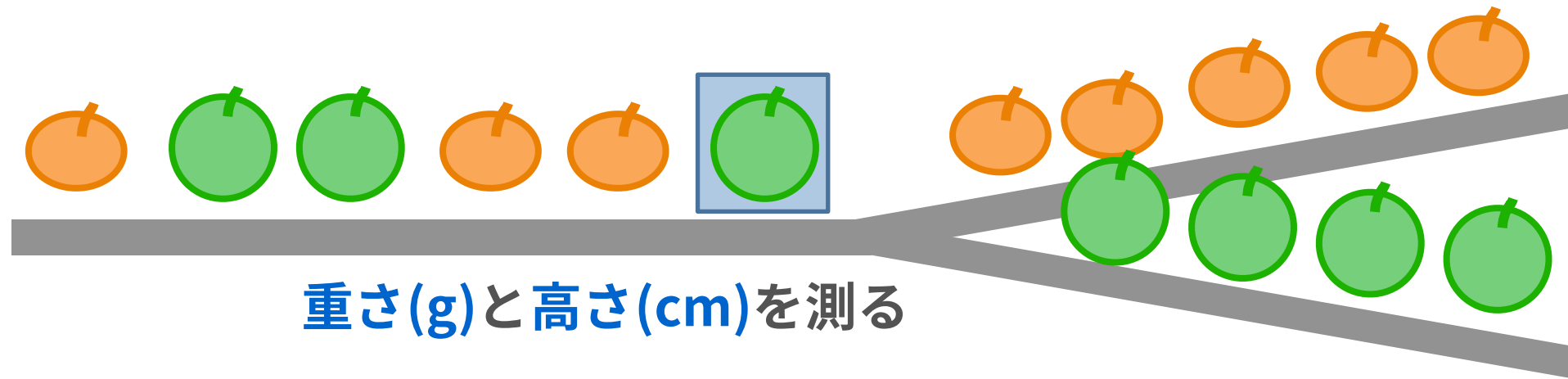


- りんご
- みかん

重さ(g)

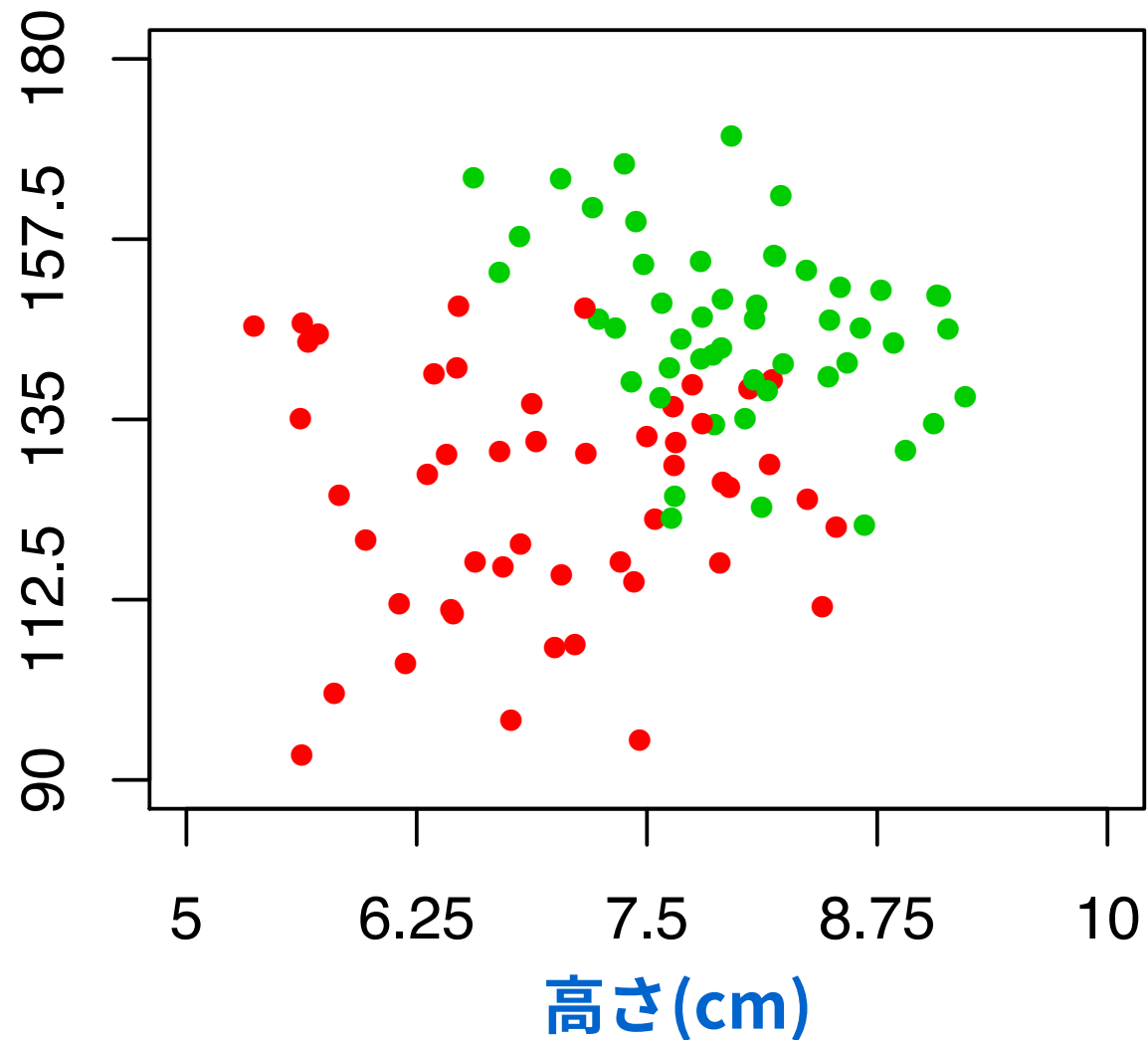


機械学習は「データを予測に変える」技術

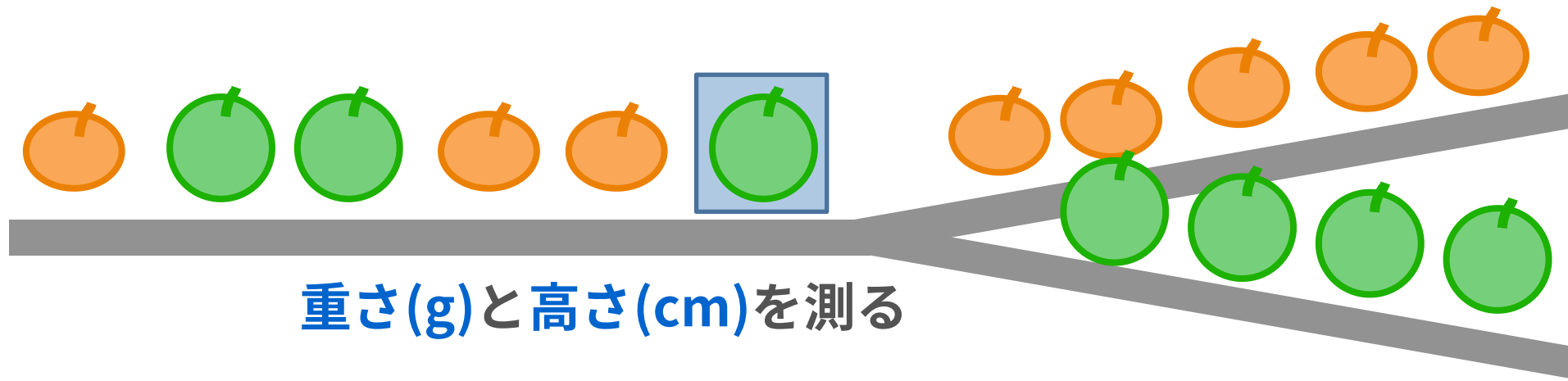


- りんご
- みかん

重さ(g)

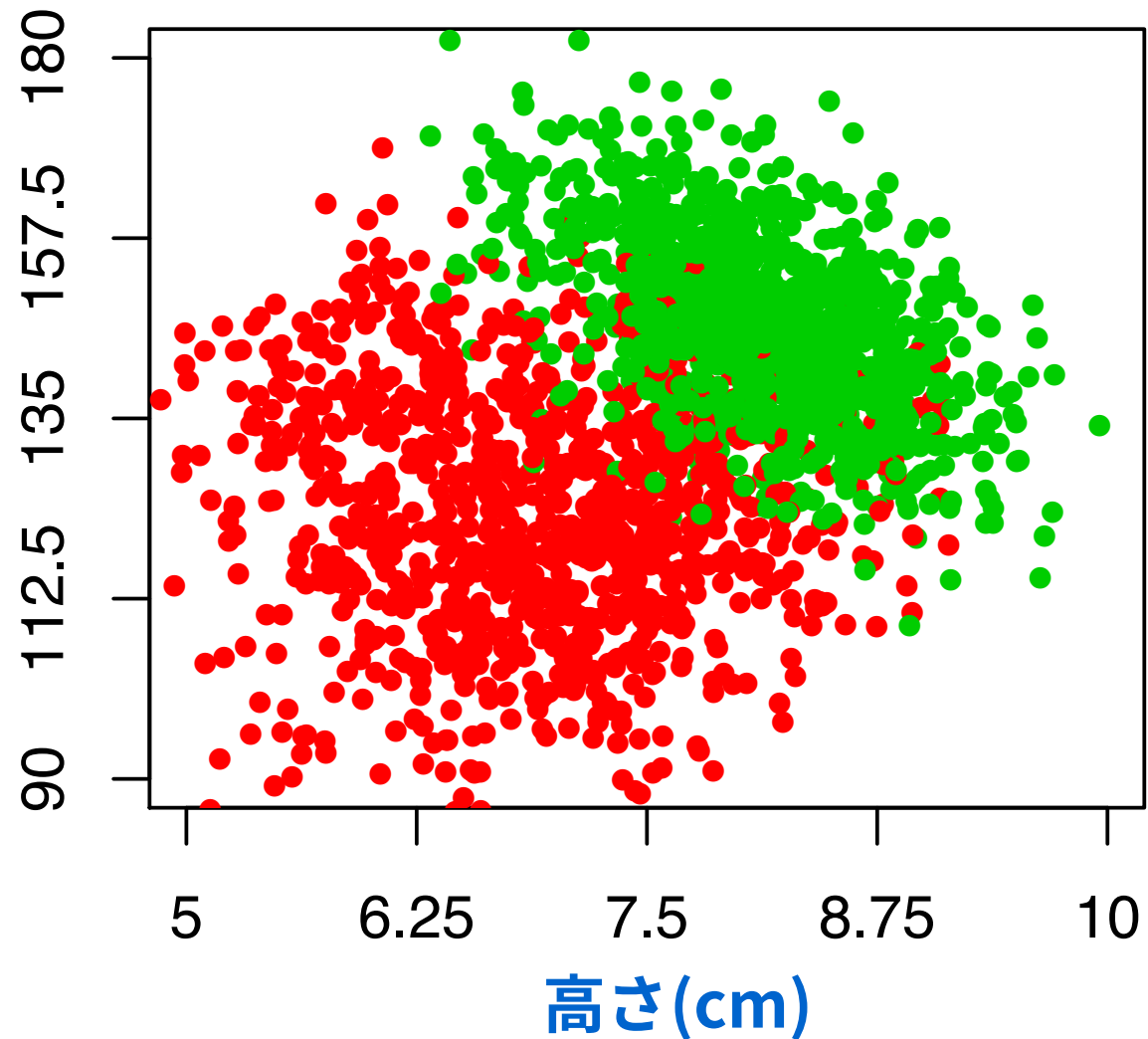


機械学習は「データを予測に変える」技術

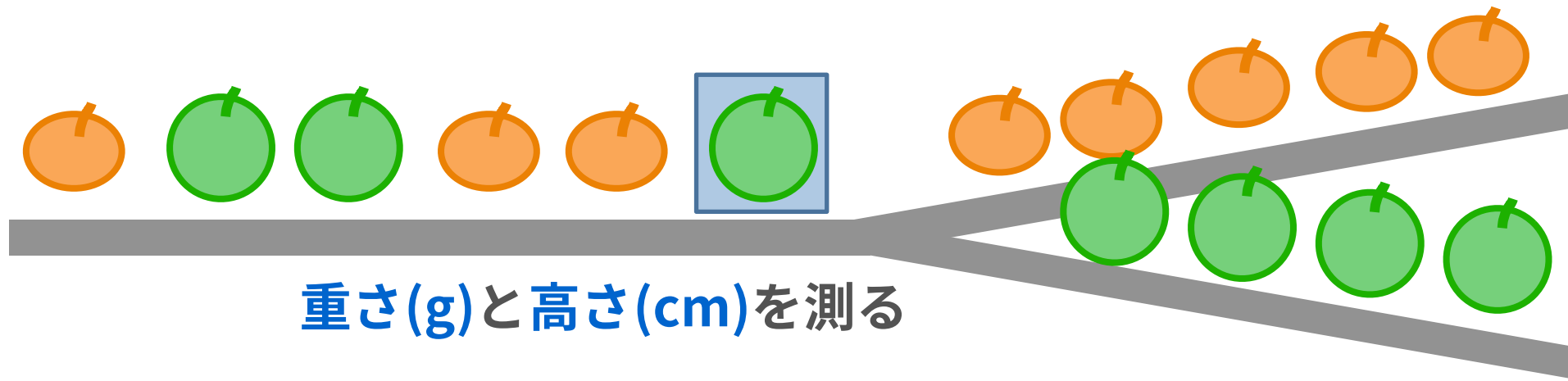


- りんご
- みかん

重さ(g)

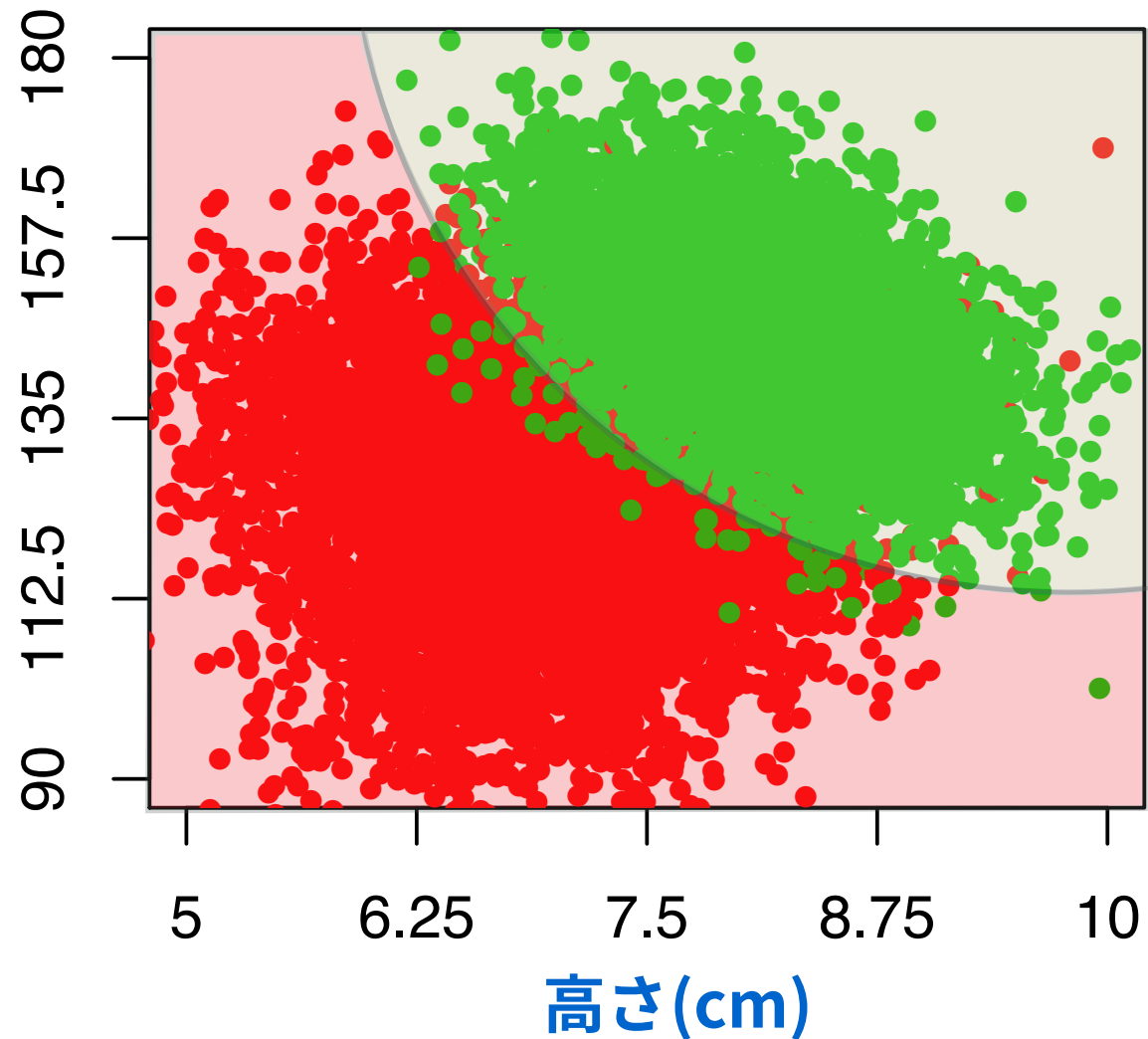


機械学習は「データを予測に変える」技術

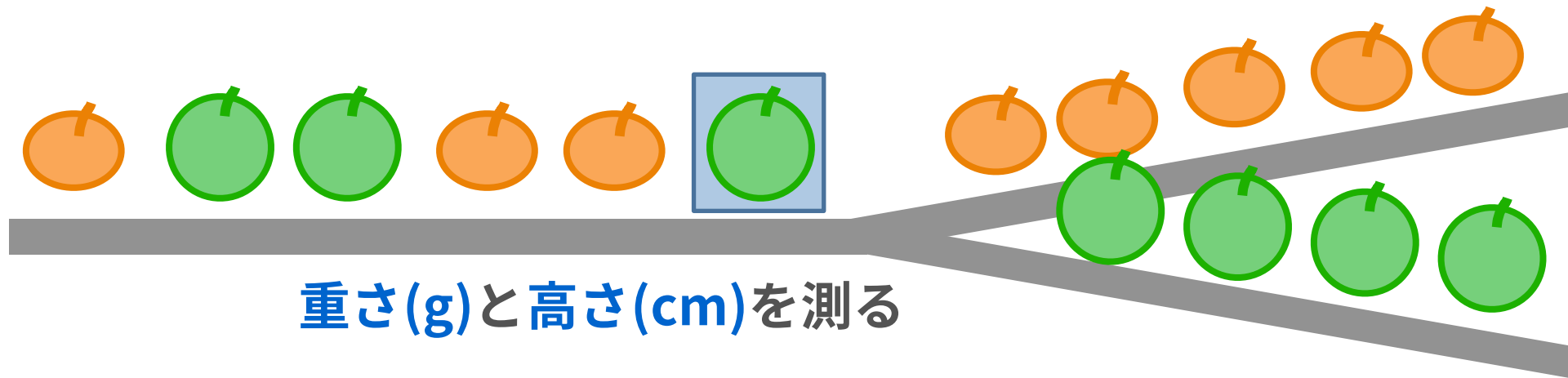


- りんご
- みかん

重さ(g)

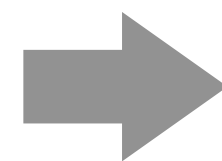
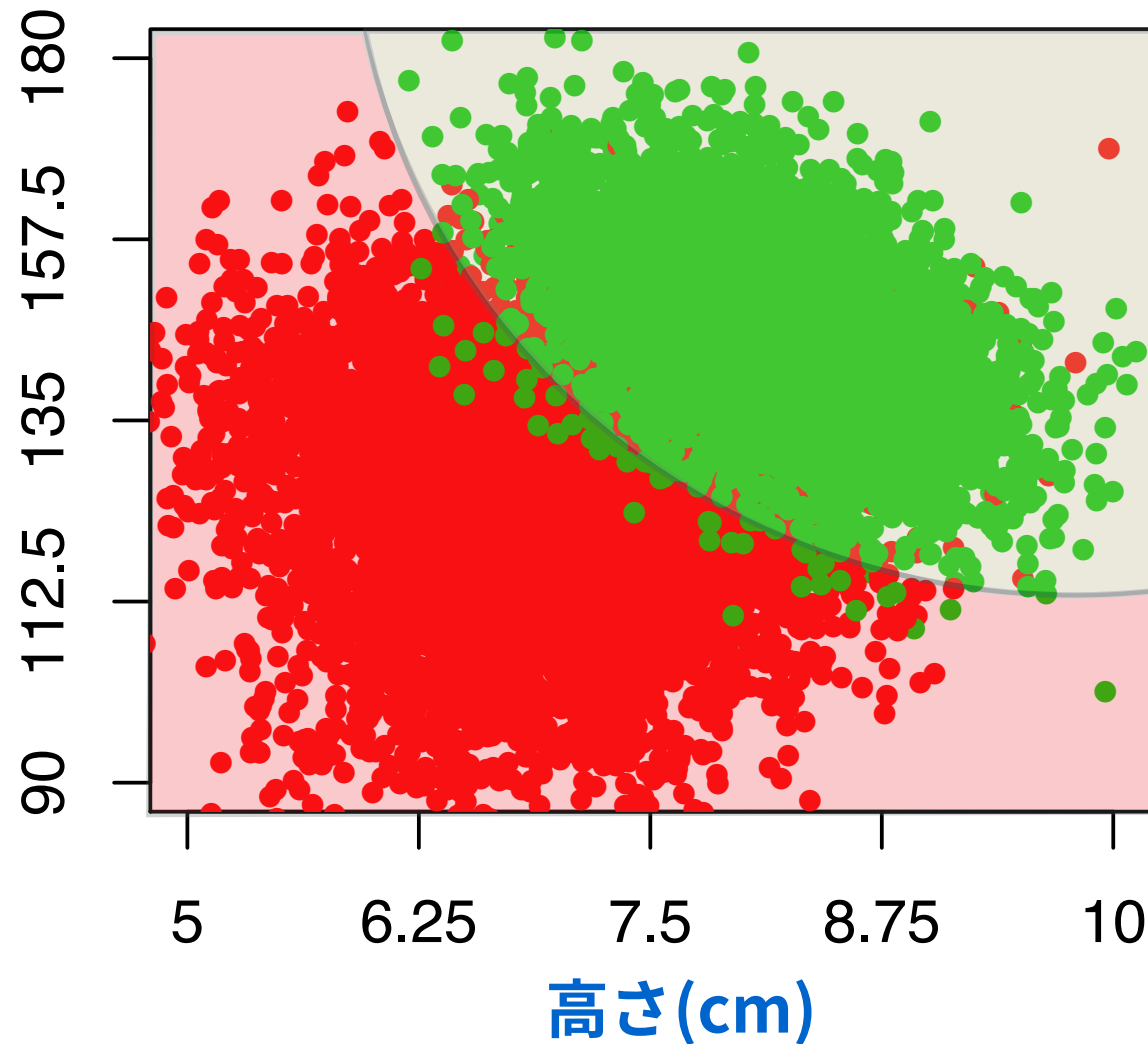


機械学習は「データを予測に変える」技術



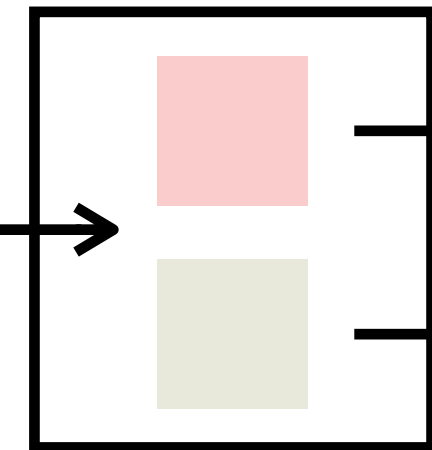
- りんご
- みかん

重さ(g)



重さ(g)
高さ(cm)

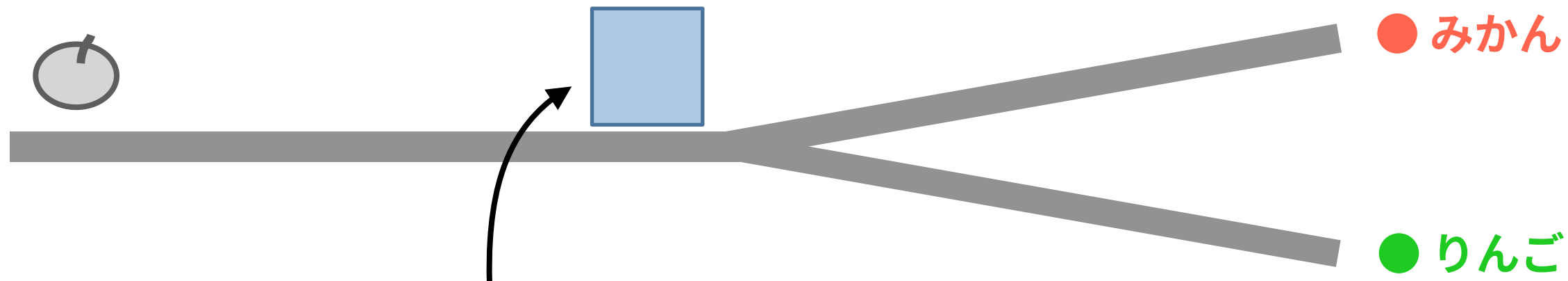
予測プログラム



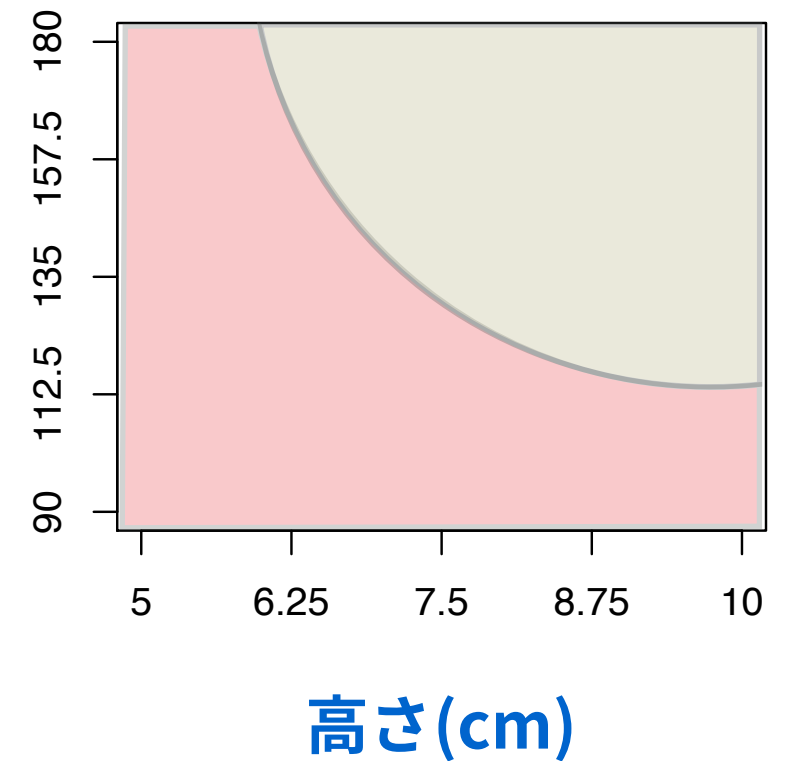
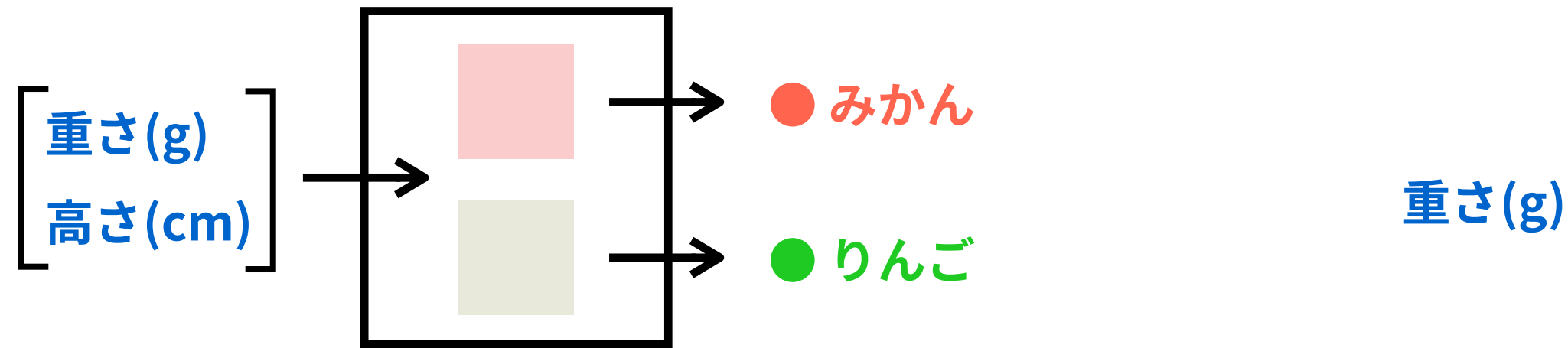
● みかん

● りんご

機械学習は「データを予測に変える」技術

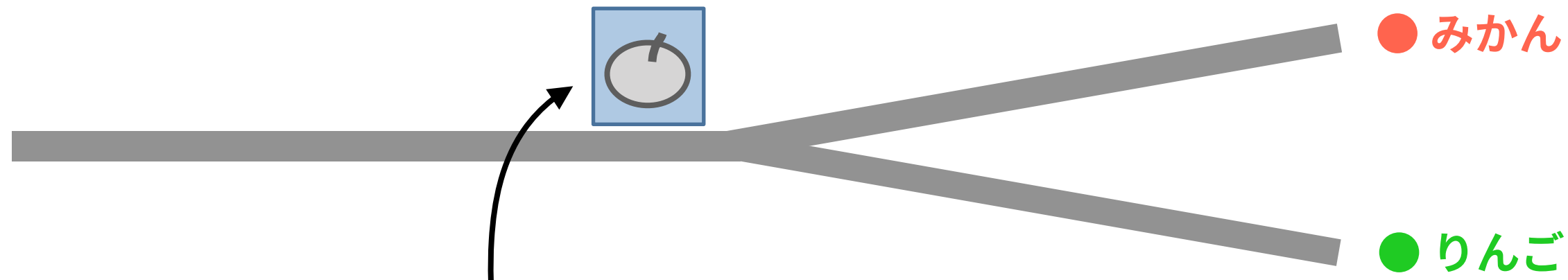


見本データから作っておいた予測プログラム

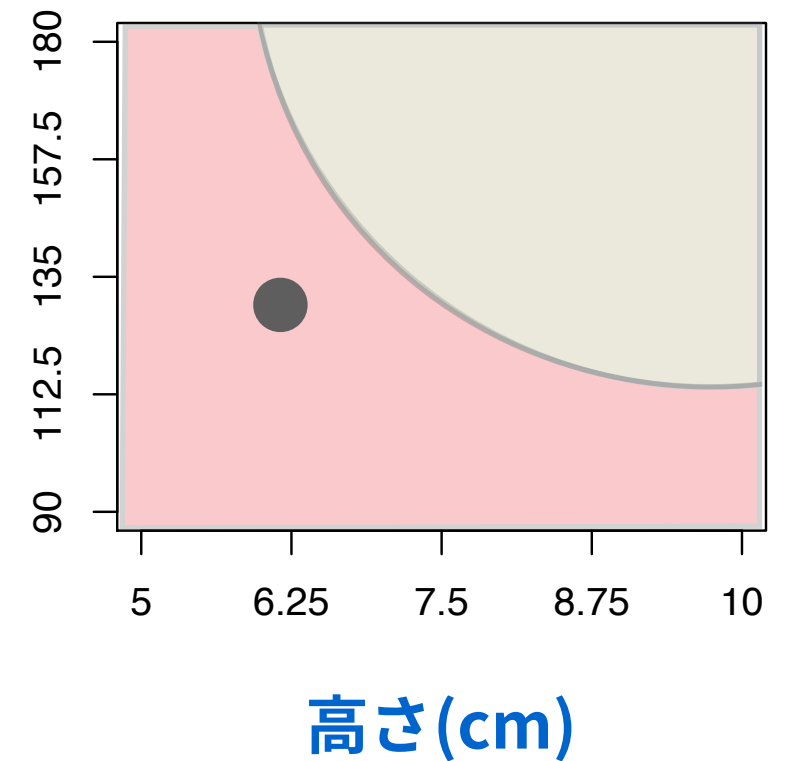
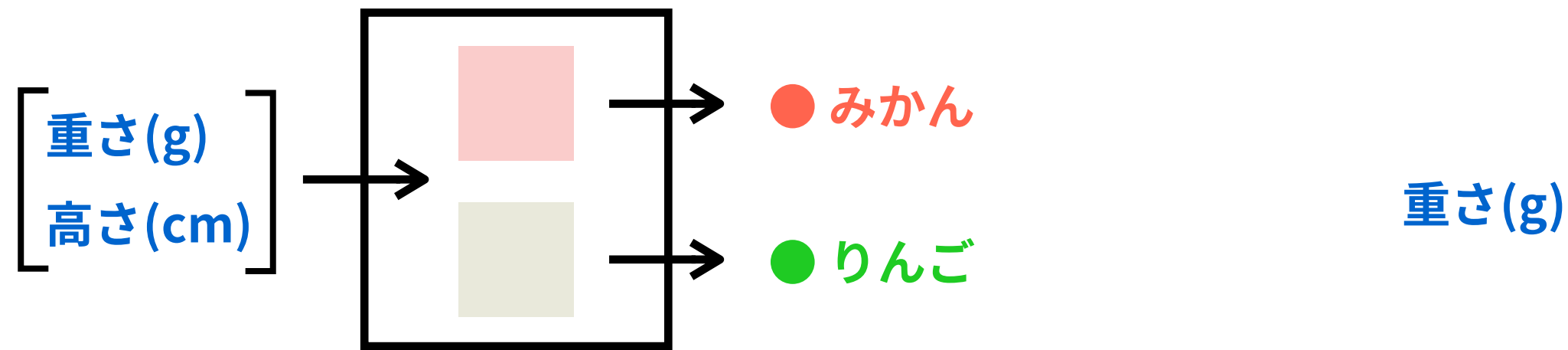


予測プログラムを作ったときに見せた見本例ではない例に対して
「みかん」 or 「りんご」を予測することができる！

機械学習は「データを予測に変える」技術

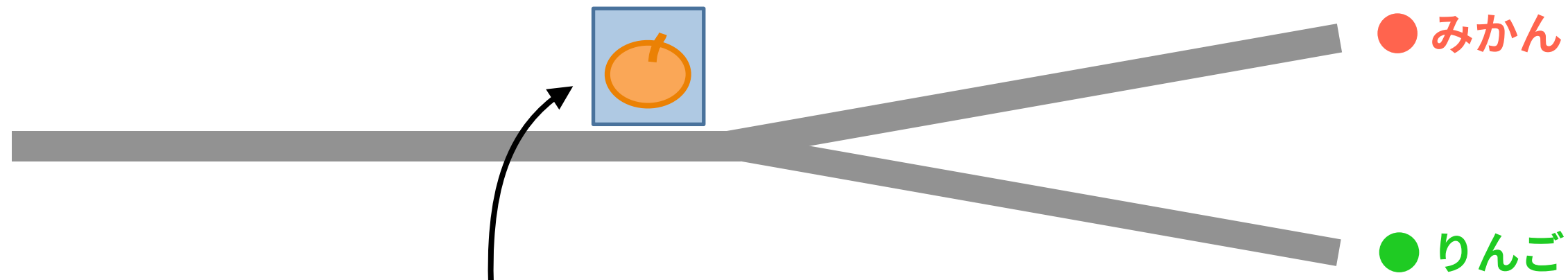


見本データから作っておいた予測プログラム

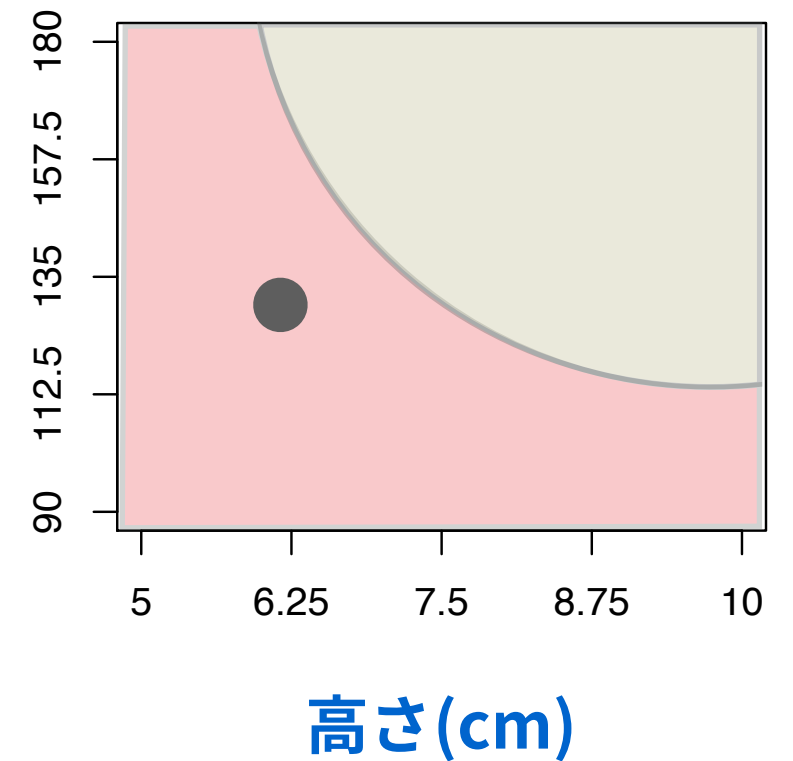
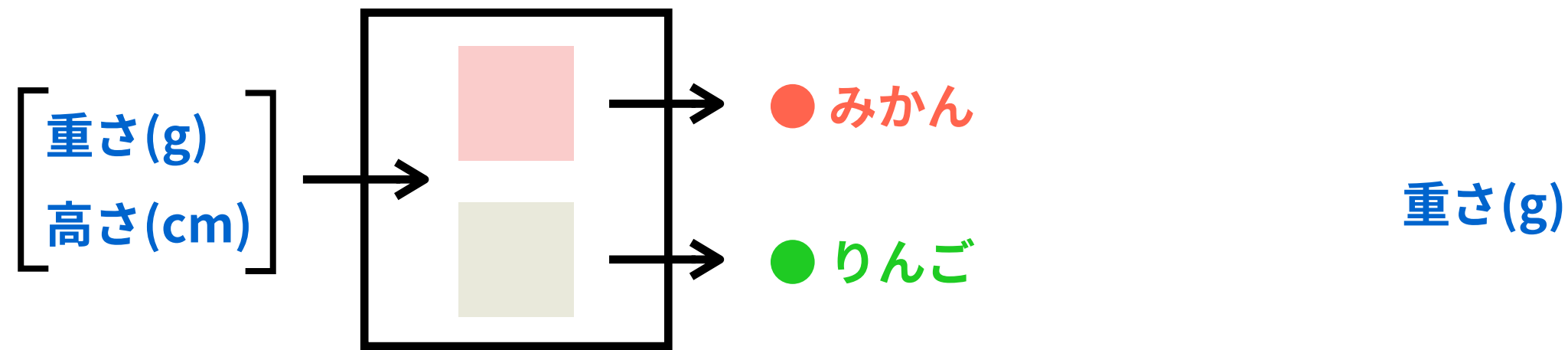


予測プログラムを作ったときに見せた見本例ではない例に対して
「みかん」 or 「りんご」を予測することができる！

機械学習は「データを予測に変える」技術

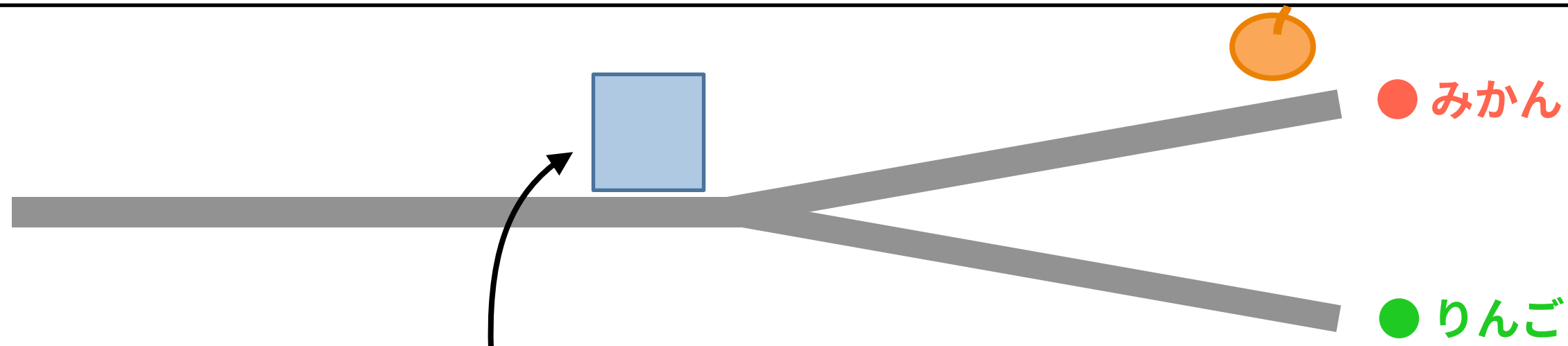


見本データから作っておいた予測プログラム

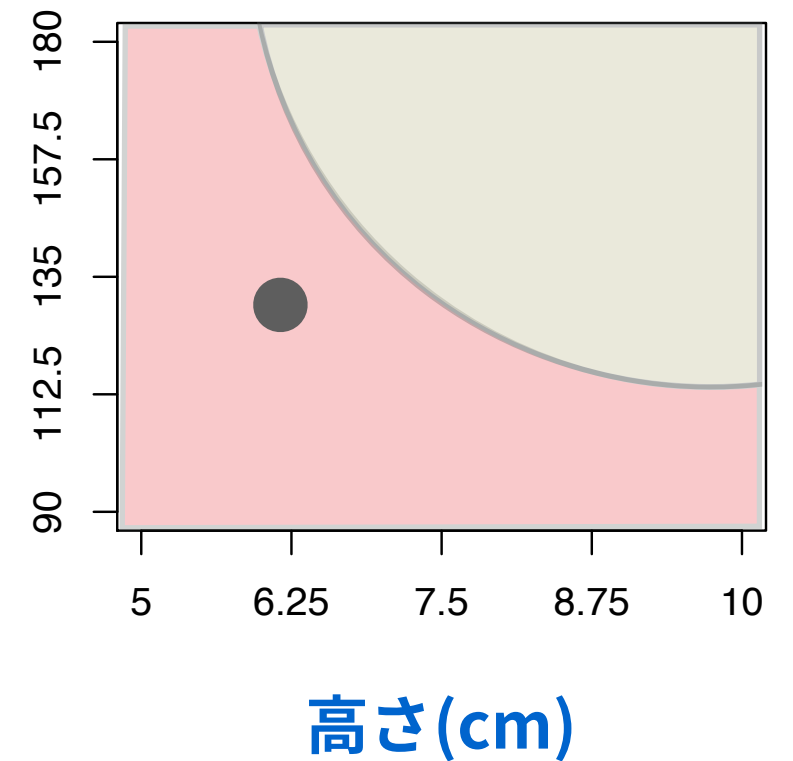
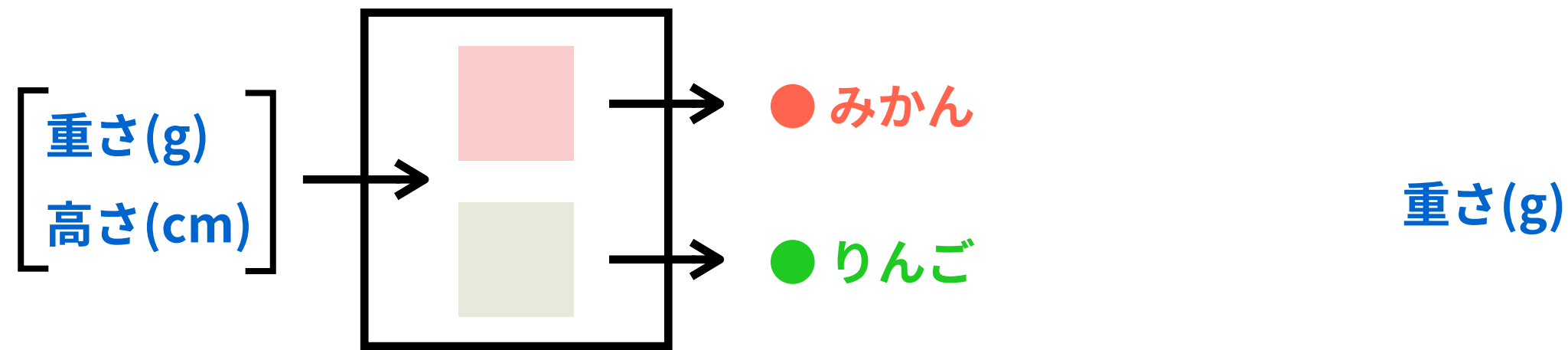


予測プログラムを作ったときに見せた見本例ではない例に対して
「みかん」 or 「りんご」を予測することができる！

機械学習は「データを予測に変える」技術

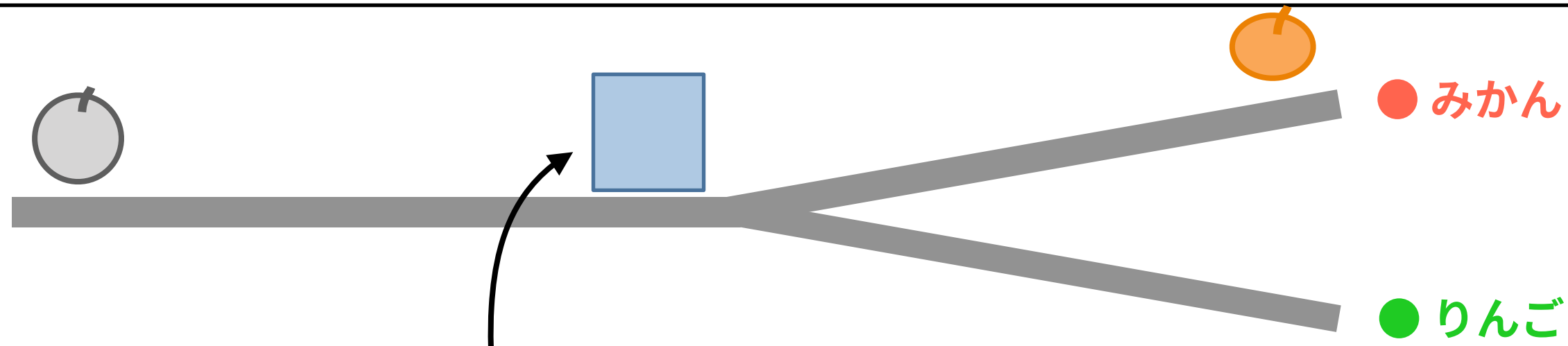


見本データから作っておいた予測プログラム

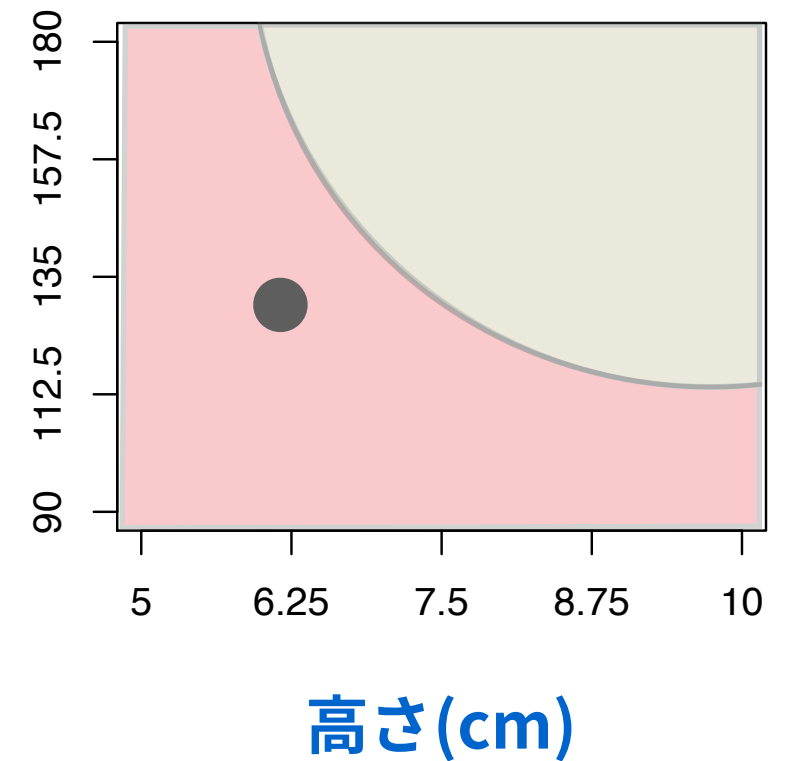
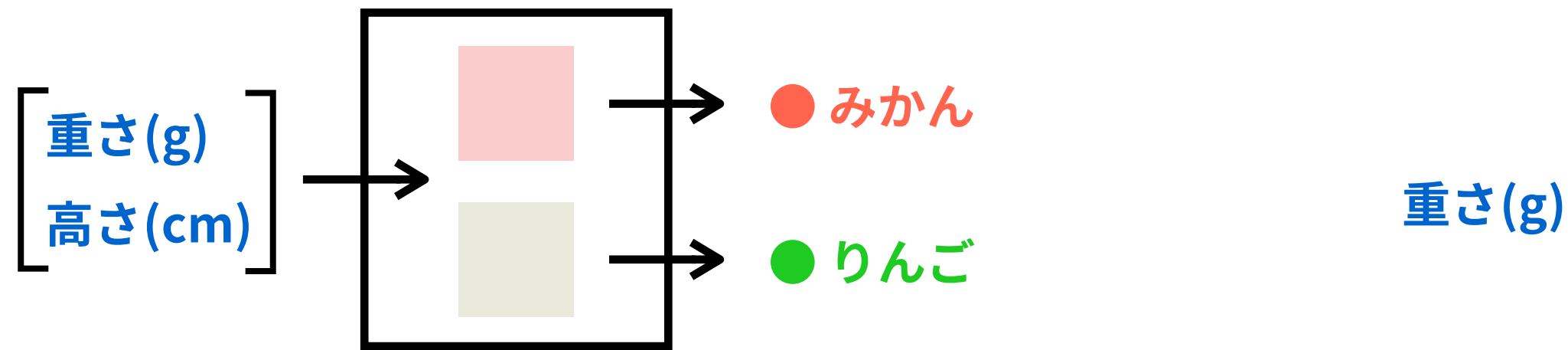


予測プログラムを作ったときに見せた見本例ではない例に対して
「みかん」 or 「りんご」を予測することができる！

機械学習は「データを予測に変える」技術

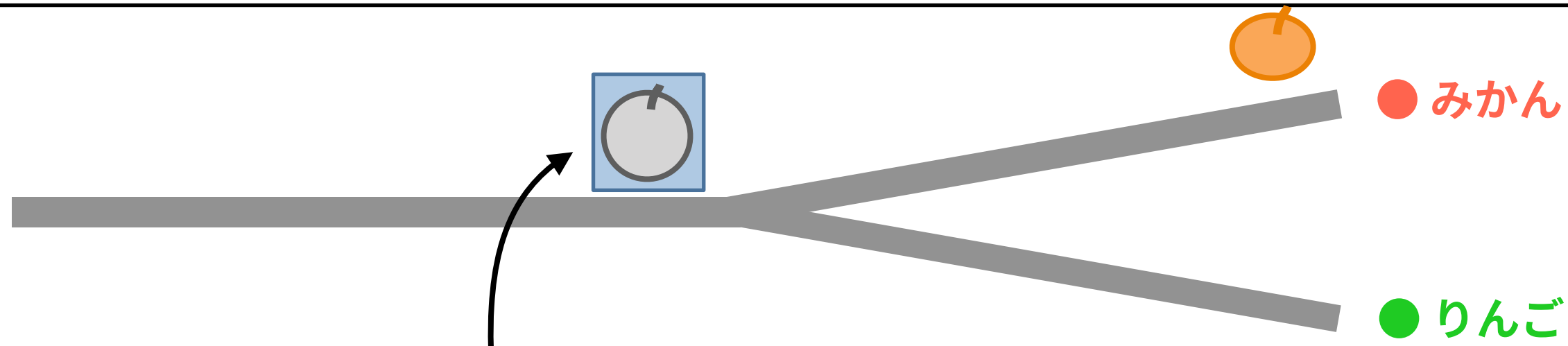


見本データから作っておいた予測プログラム

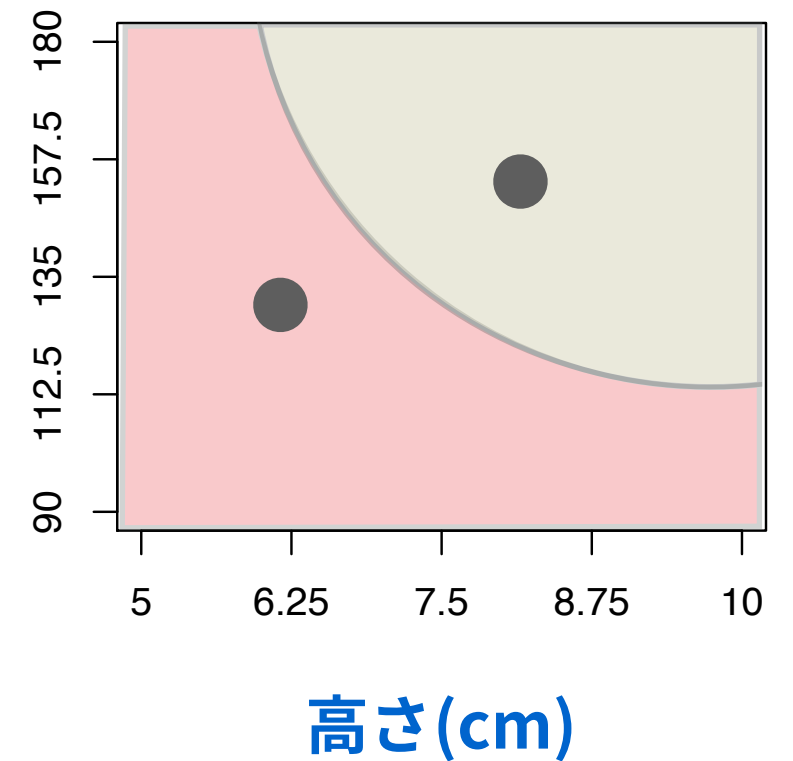
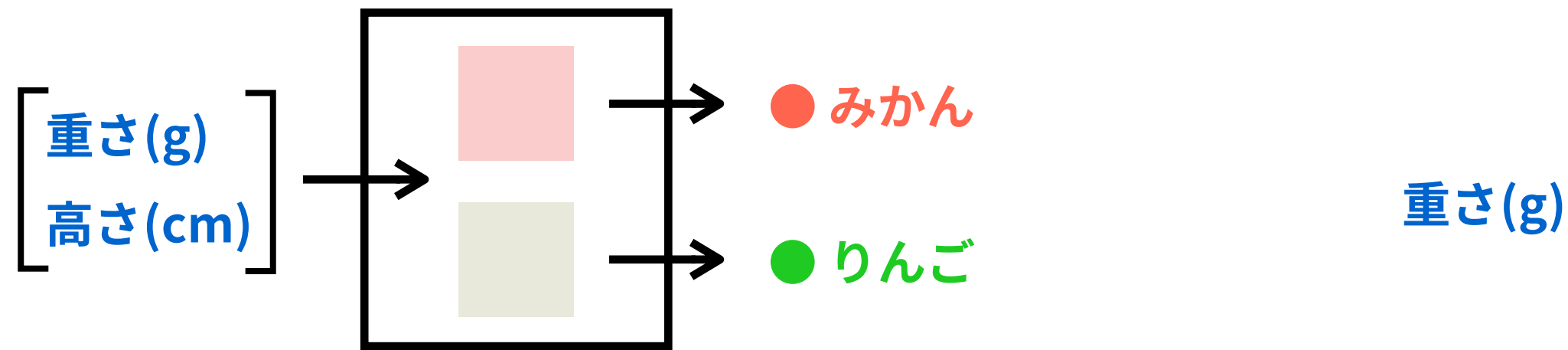


予測プログラムを作ったときに見せた見本例ではない例に対して
「みかん」 or 「りんご」を予測することができる！

機械学習は「データを予測に変える」技術

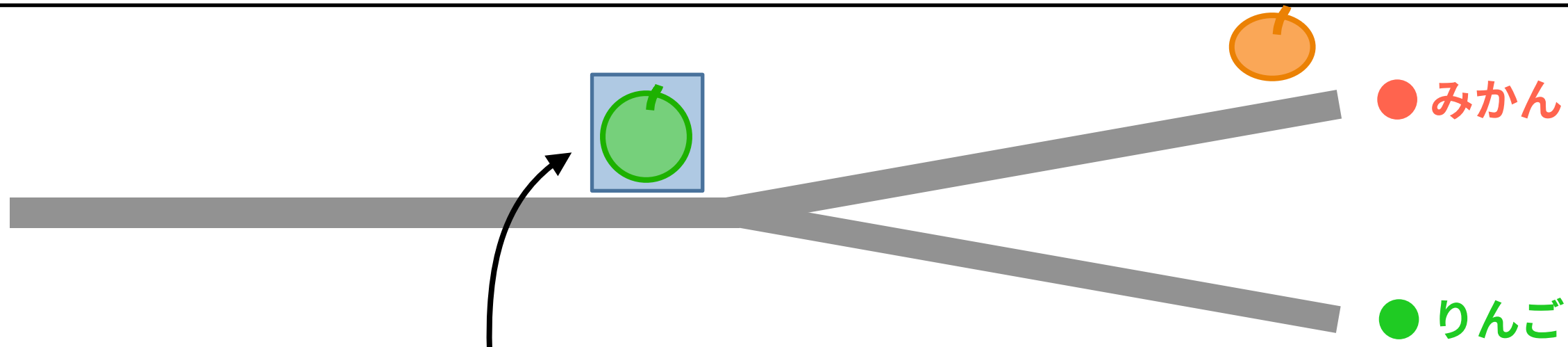


見本データから作っておいた予測プログラム

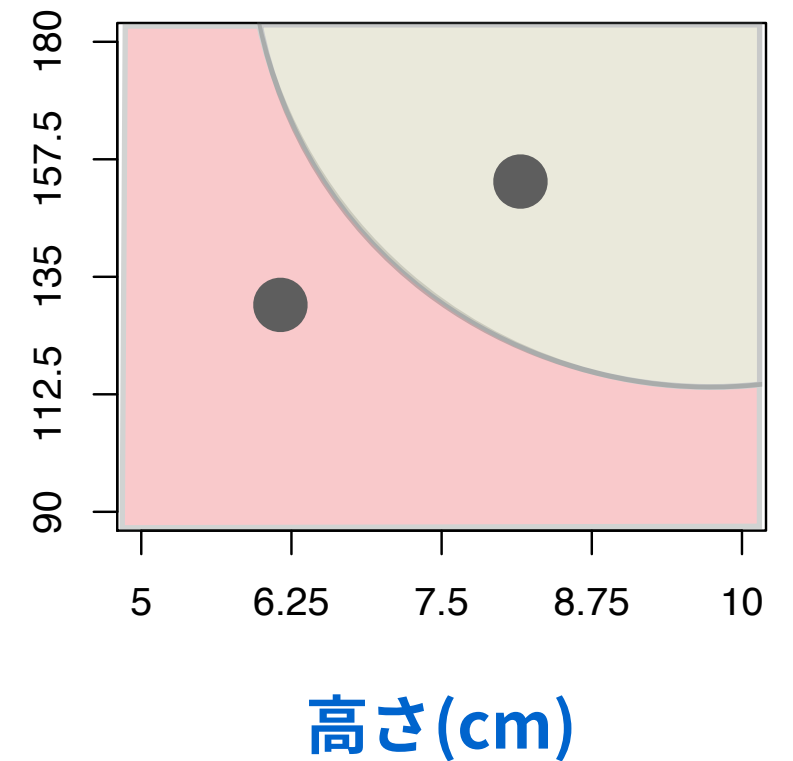
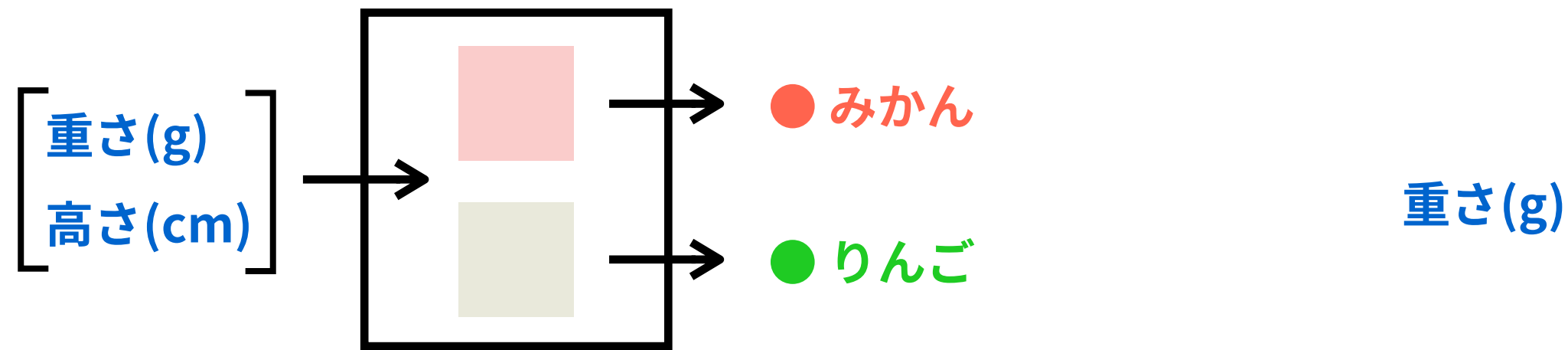


予測プログラムを作ったときに見せた見本例ではない例に対して
「みかん」 or 「りんご」を予測することができる！

機械学習は「データを予測に変える」技術

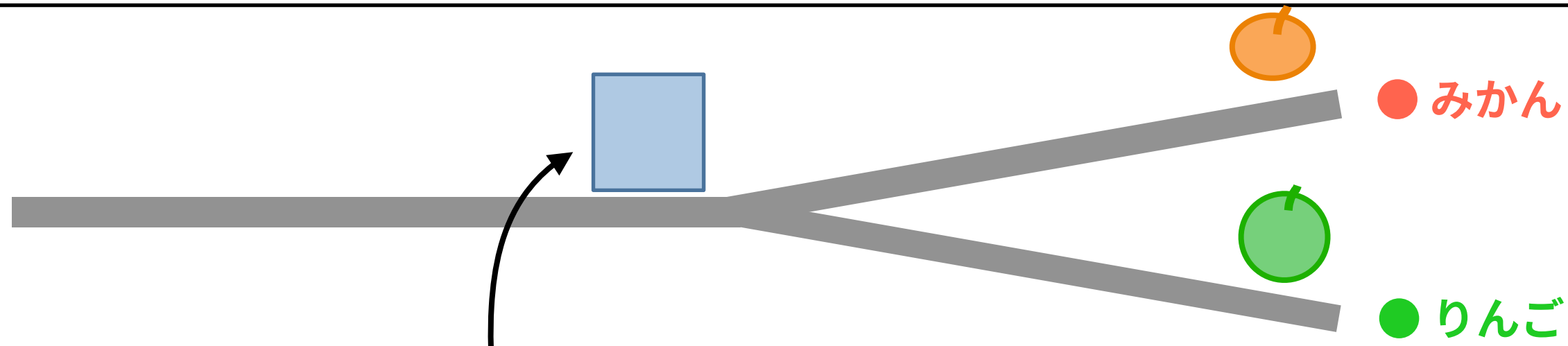


見本データから作っておいた予測プログラム

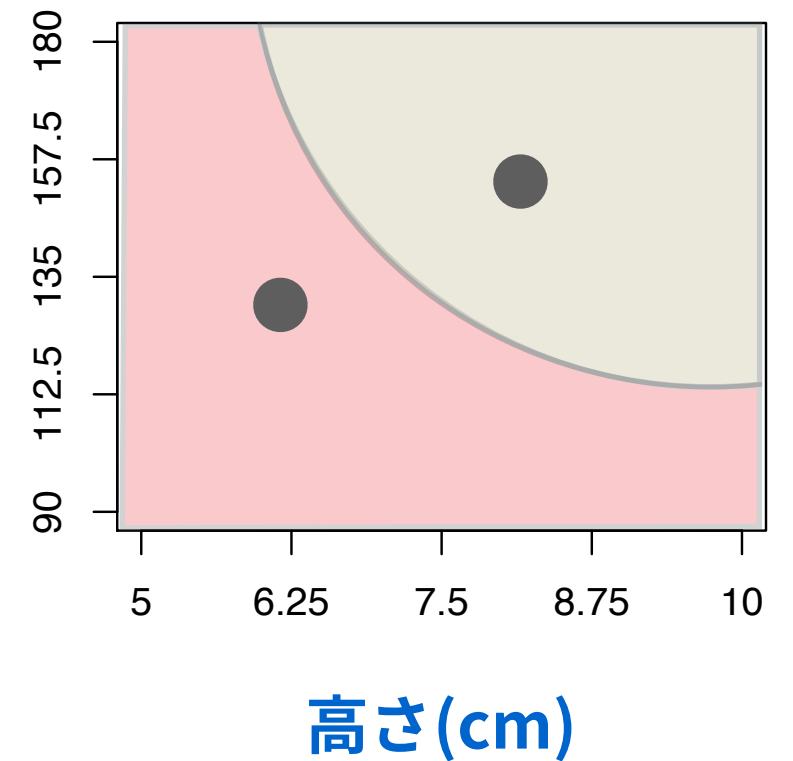
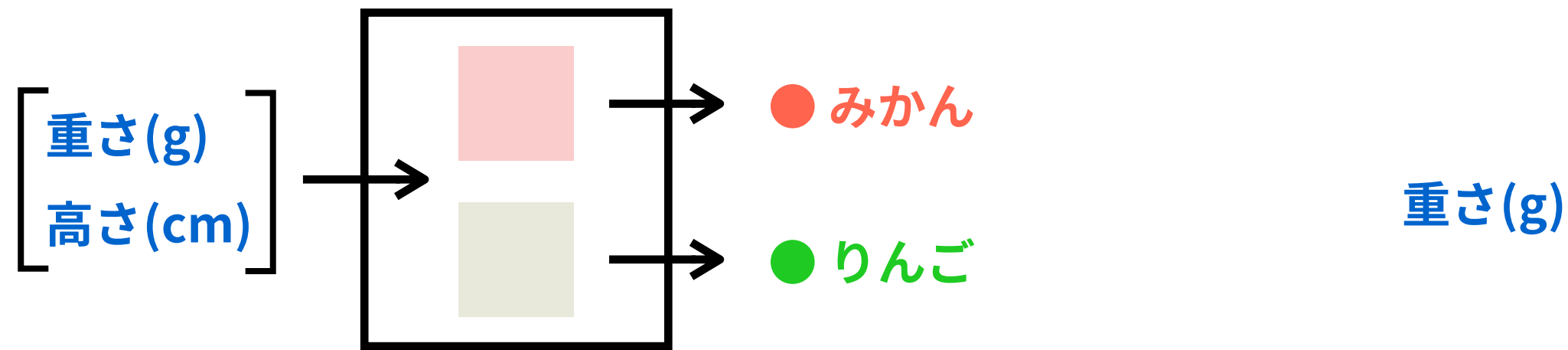


予測プログラムを作ったときに見せた見本例ではない例に対して
「みかん」 or 「りんご」を予測することができる！

機械学習は「データを予測に変える」技術



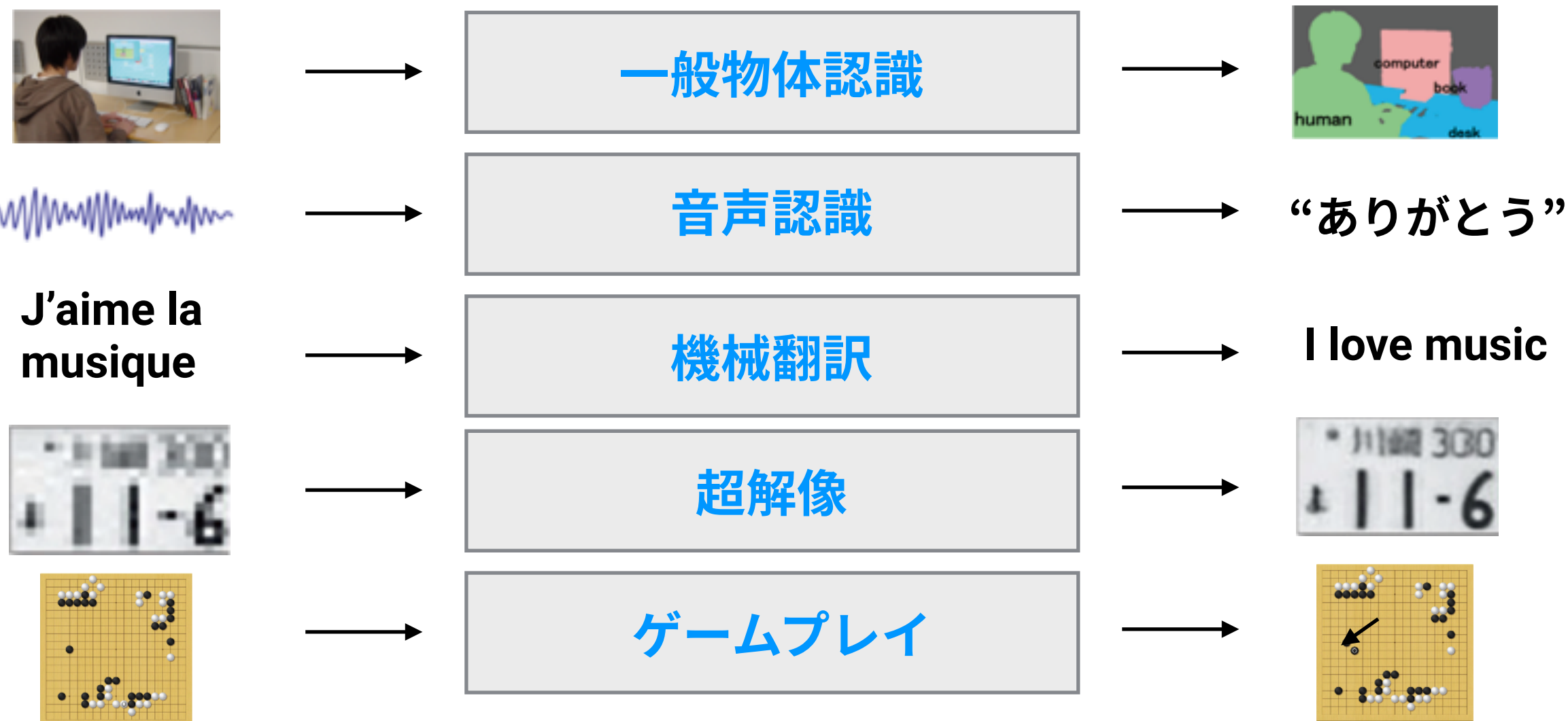
見本データから作っておいた予測プログラム



予測プログラムを作ったときに見せた見本例ではない例に対して
「みかん」 or 「りんご」を予測することができる！

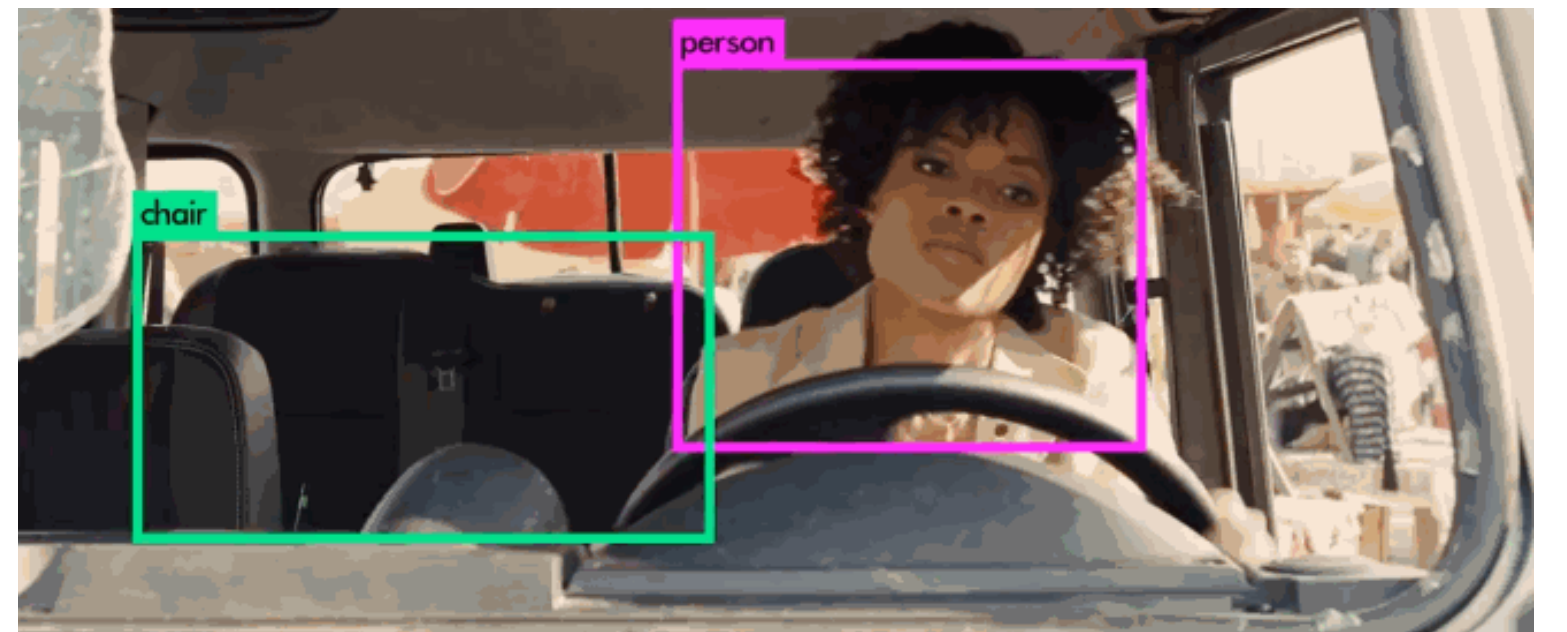
機械学習は「新しい(雑な)コンピュータプログラムの作り方」

プログラムの入力と出力の関係がよく分からない場合でも、
たくさんの入出力の**見本データ**によって間接的にそれを再現できるプログラムを作り出す技術



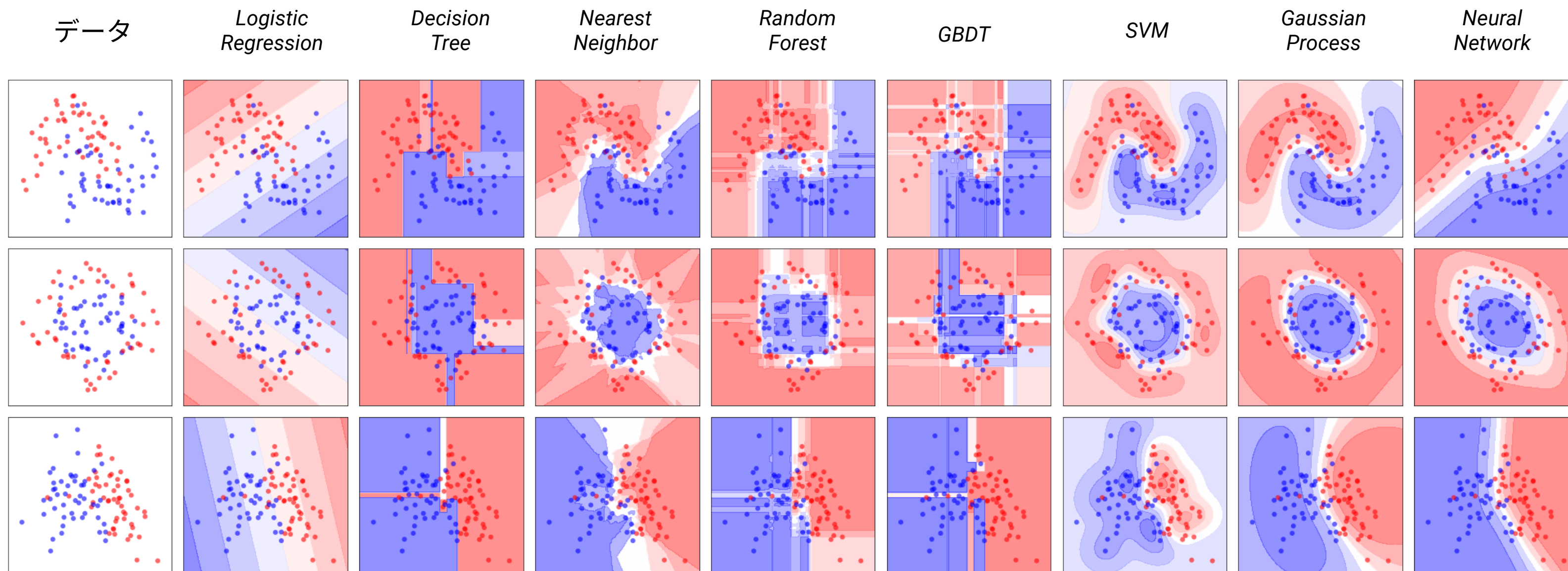
機械学習は「新しい(雑な)コンピュータプログラムの作り方」

この単純なしくみは上手に使うと「めちやくちゃ強力」でいろいろな楽しいこともできる！



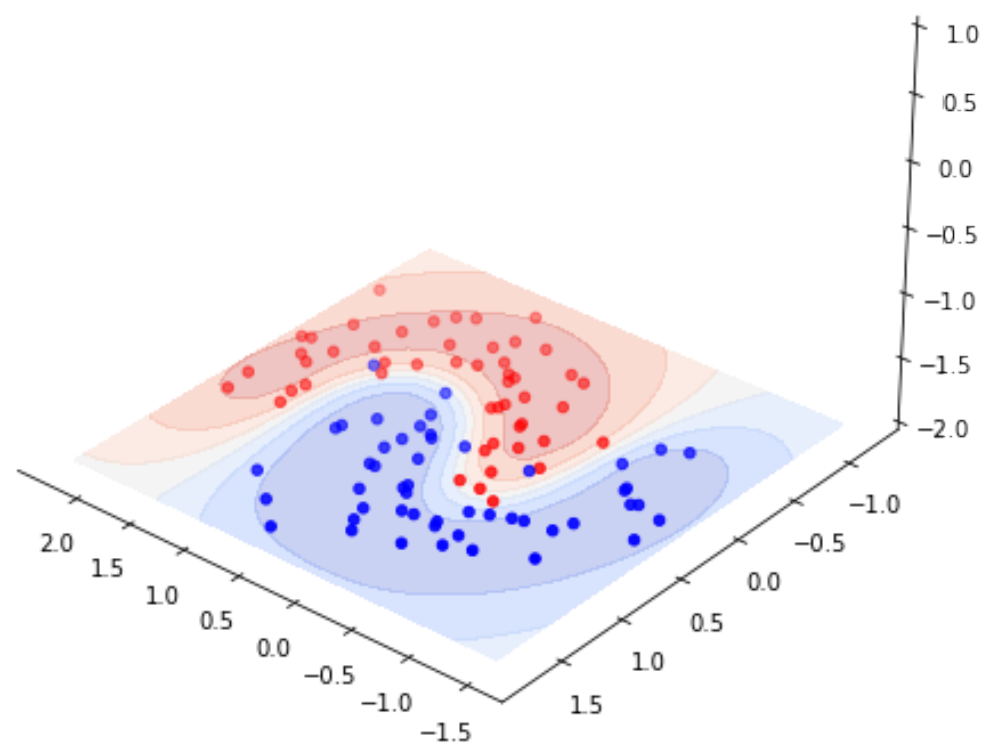
機械学習は「新しい(雑な)コンピュータプログラムの作り方」

機械学習のアルゴリズムはたくさんある。違いは境界線の引き方の方針



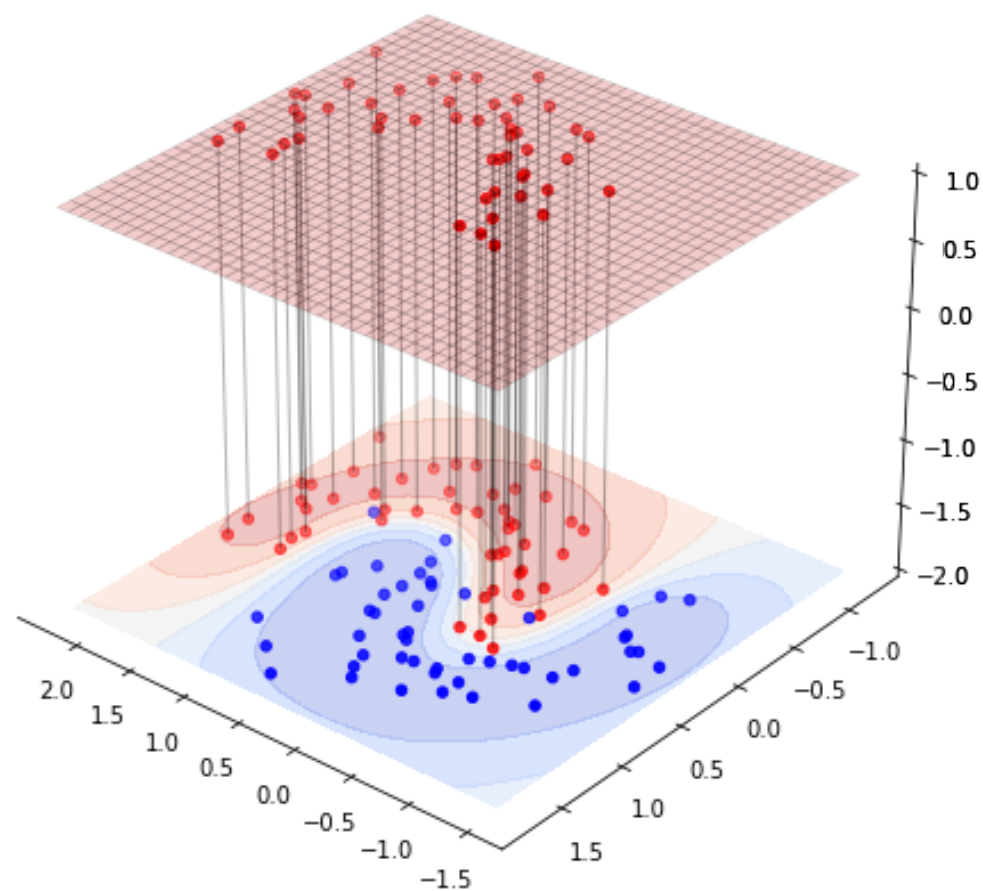
仕組みは曲面フィッティングによるデータ内挿

内部原理は「**曲面モデル**」を点にフィッティングしているだけ



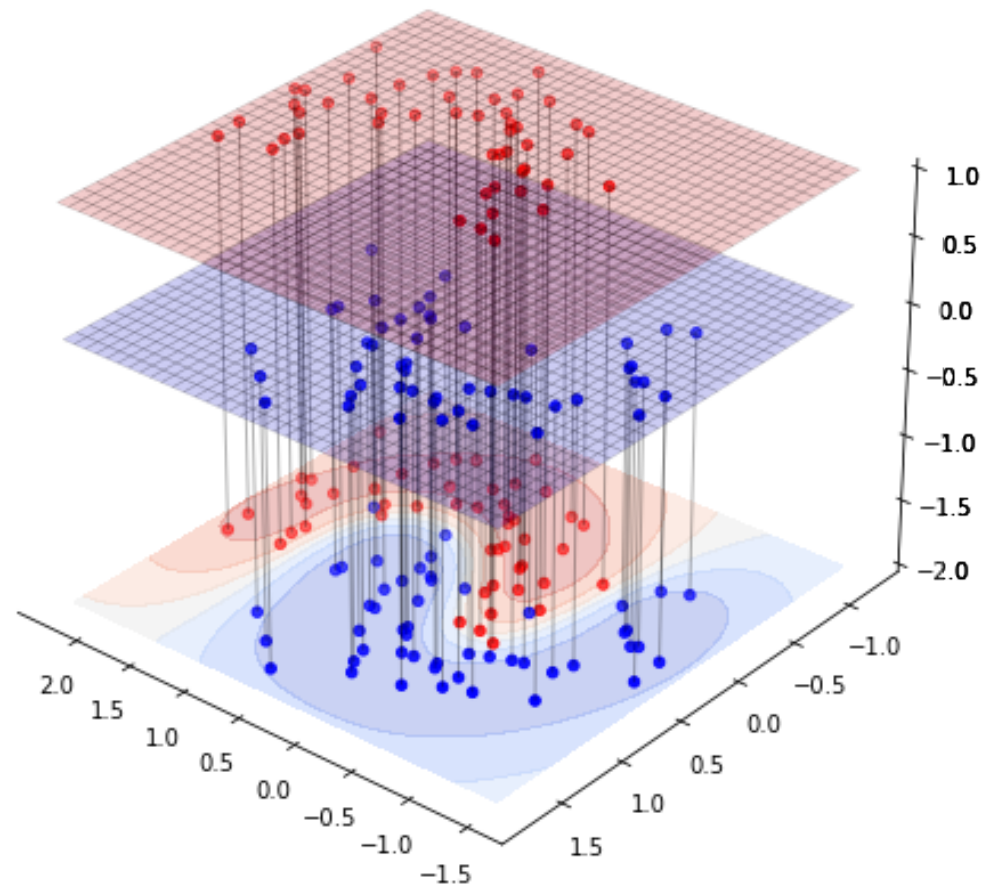
仕組みは曲面フィッティングによるデータ内挿

内部原理は「**曲面モデル**」を点にフィッティングしているだけ



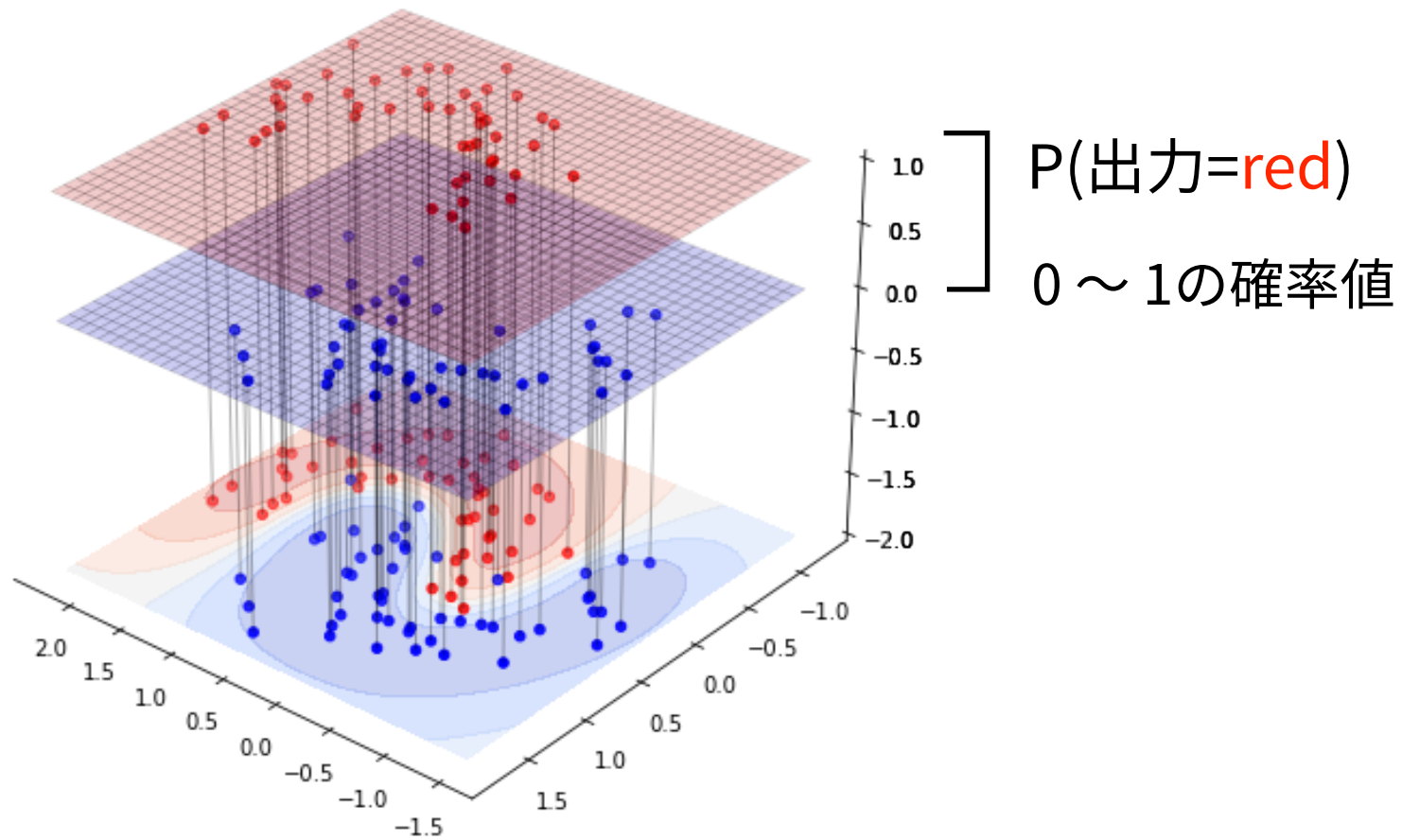
仕組みは曲面フィッティングによるデータ内挿

内部原理は「**曲面モデル**」を点にフィッティングしているだけ



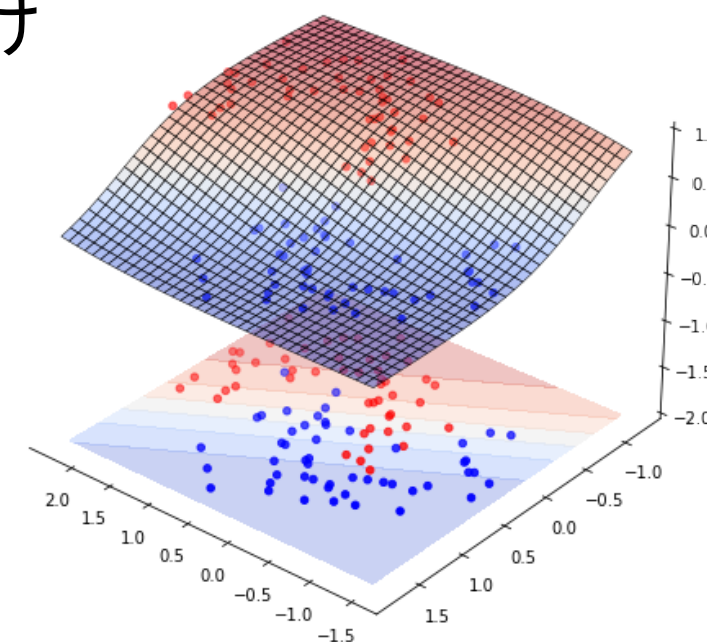
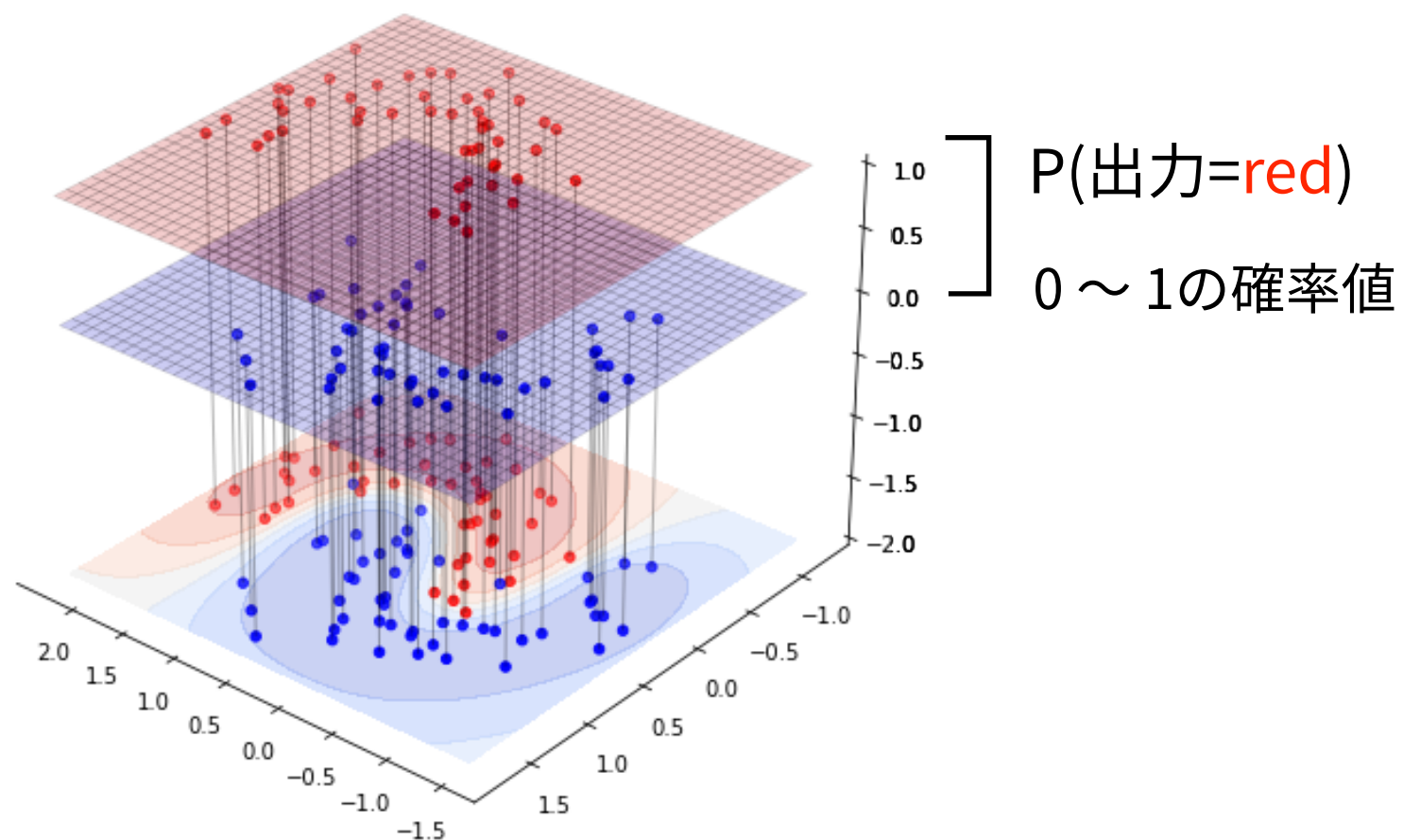
仕組みは曲面フィッティングによるデータ内挿

内部原理は「**曲面モデル**」を点にフィッティングしているだけ

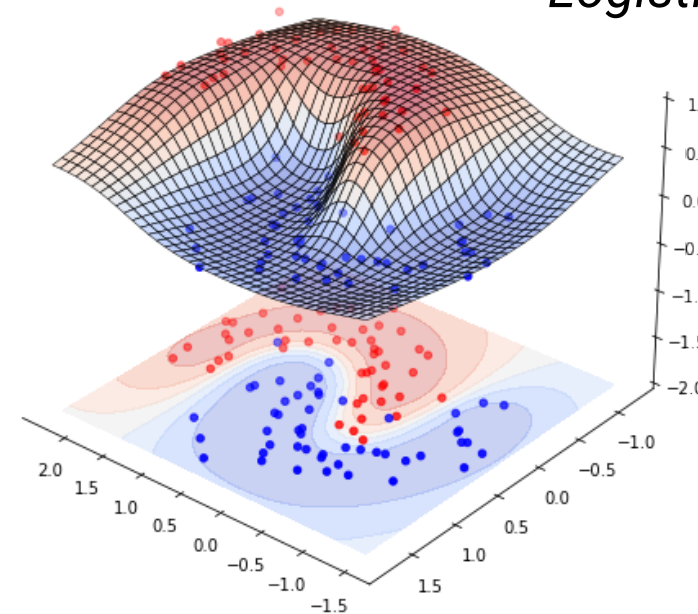


仕組みは曲面フィッティングによるデータ内挿

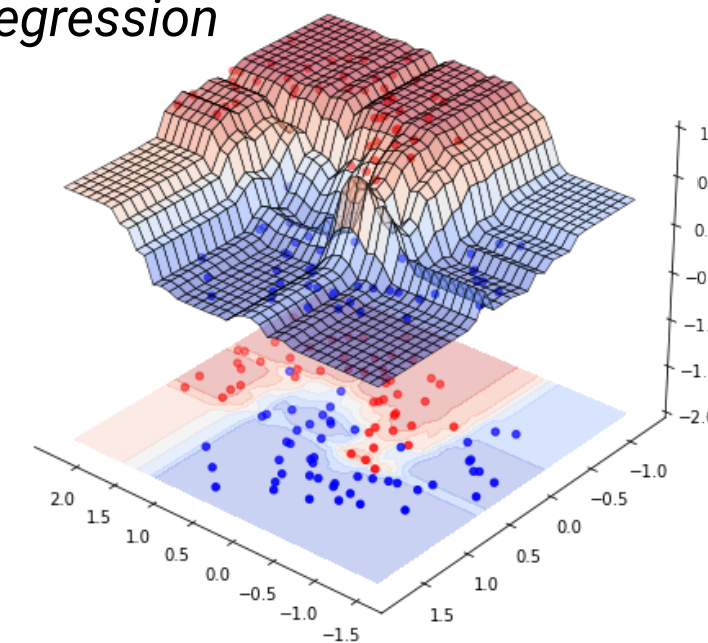
内部原理は「曲面モデル」を点にフィッティングしているだけ



Logistic Regression



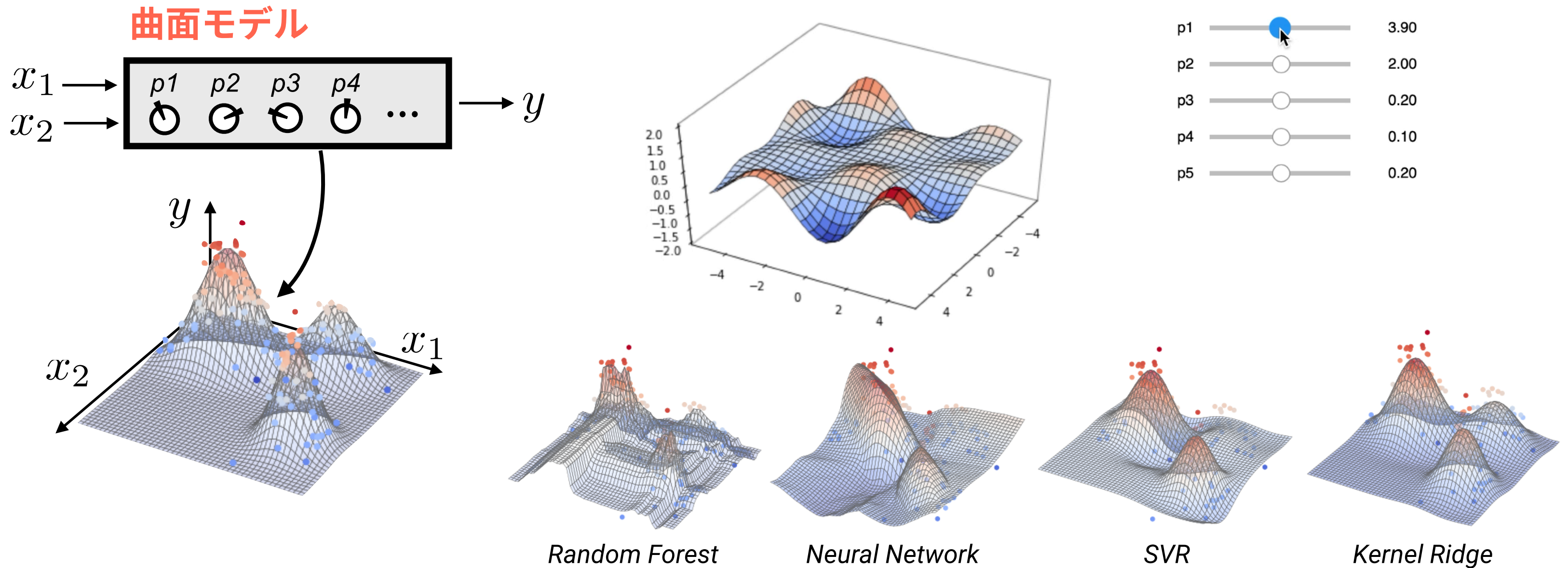
Gaussian Process



Random Forest

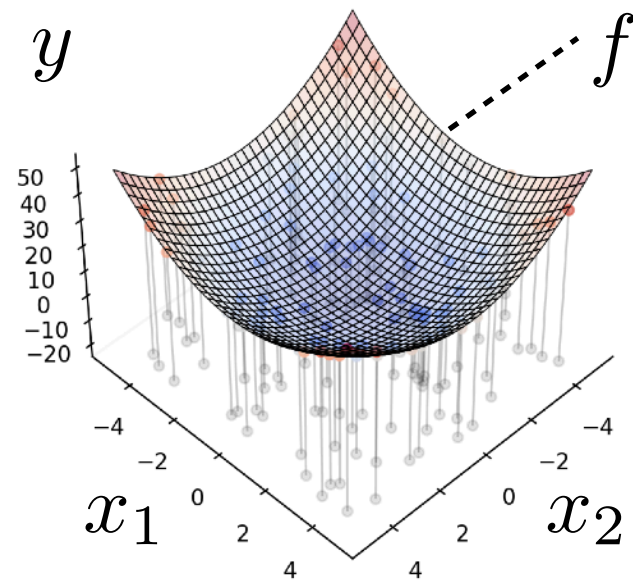
仕組みは曲面フィッティングによるデータ内挿

「曲面モデル」の内部パラメタ値を調整して見本点にあうようフィッティングする

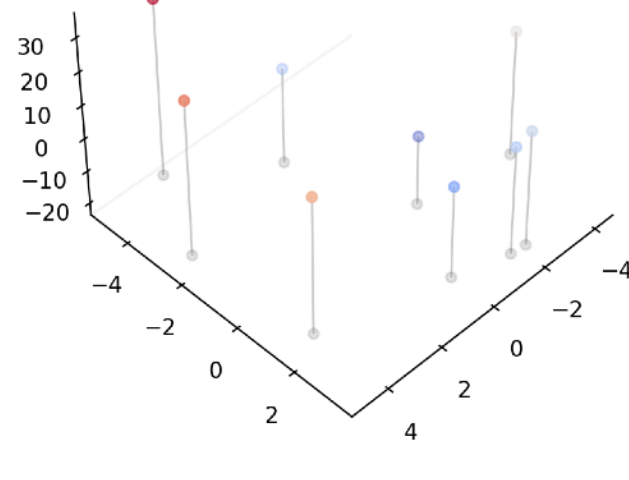


機械学習を使うのにどれくらいデータが必要ですか？

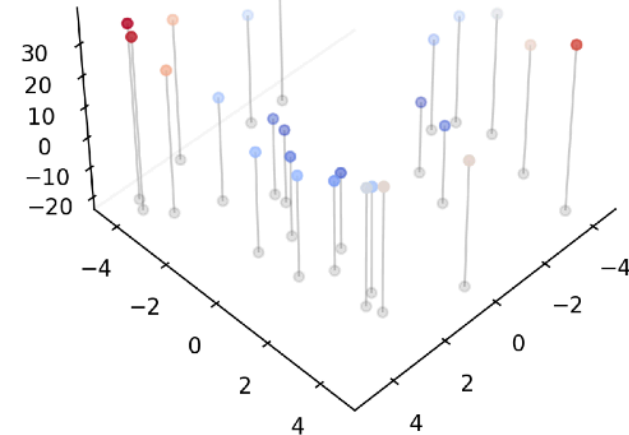
= 曲面の概形を知るのに見本点が何点必要ですか？



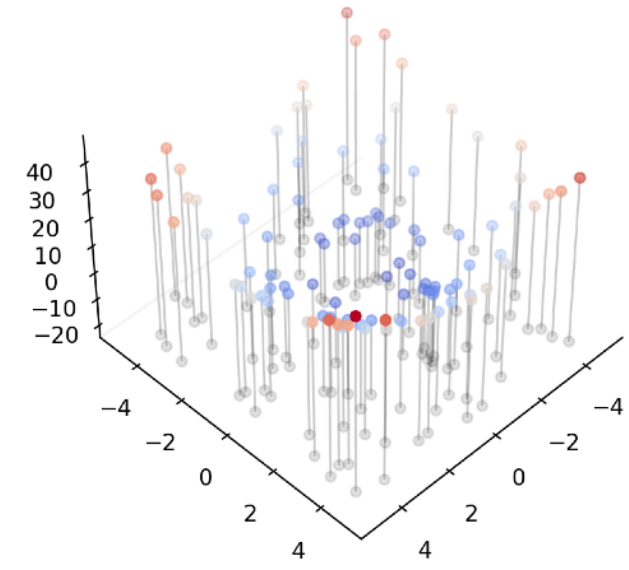
$N = 3^2 = 9$



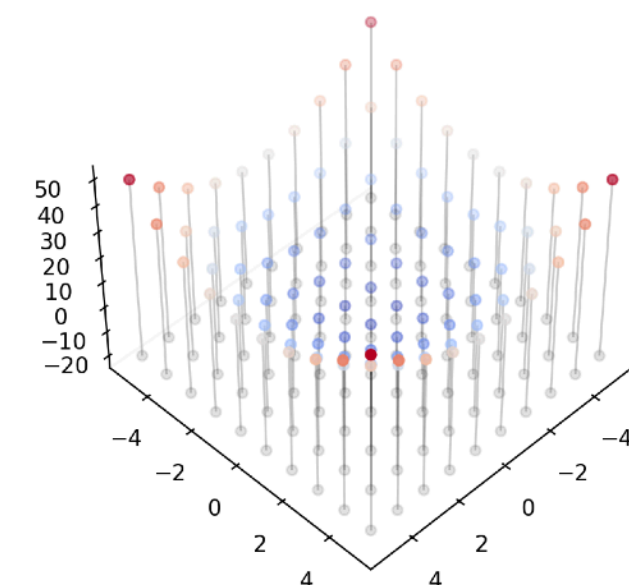
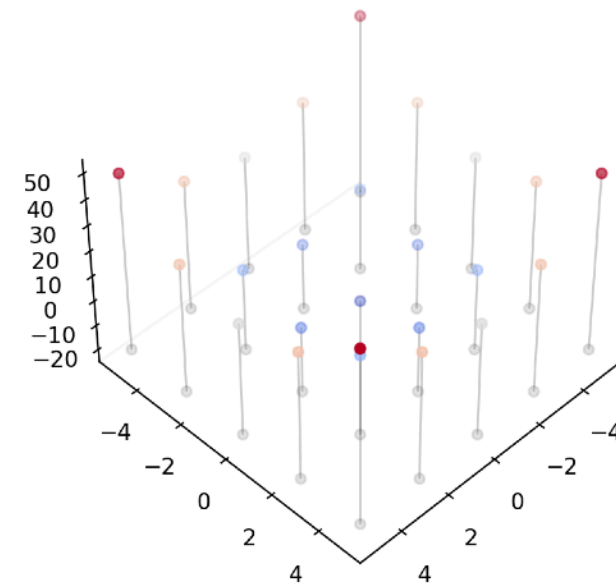
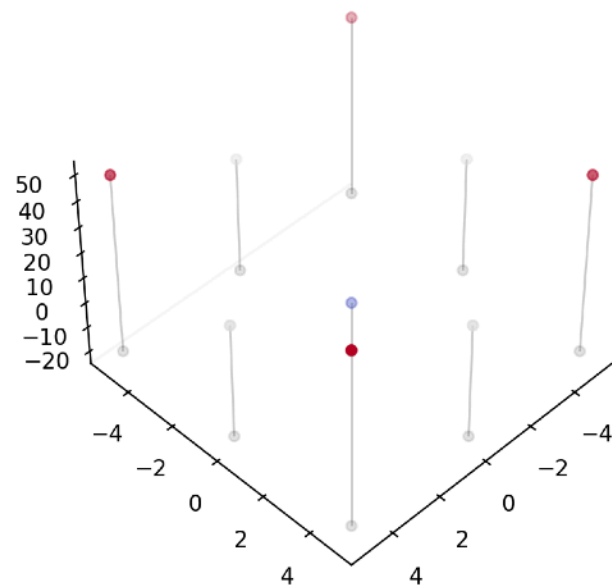
$N = 5^2 = 25$



$N = 10^2 = 100$

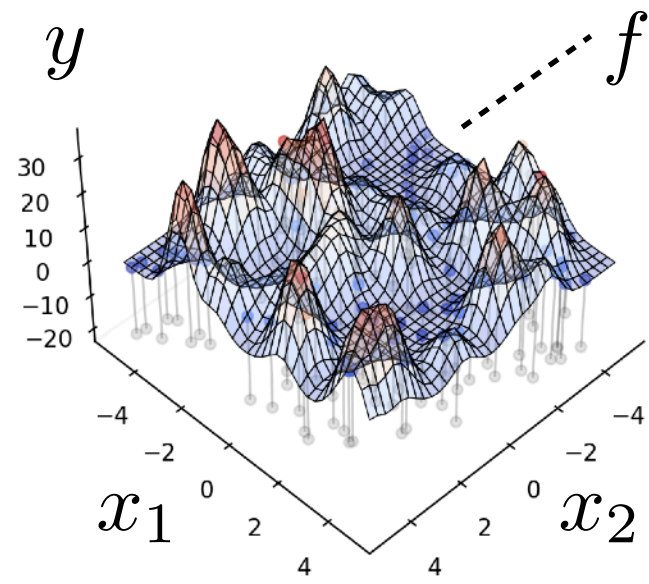


多変数(高次元)だと
データがもっと必要



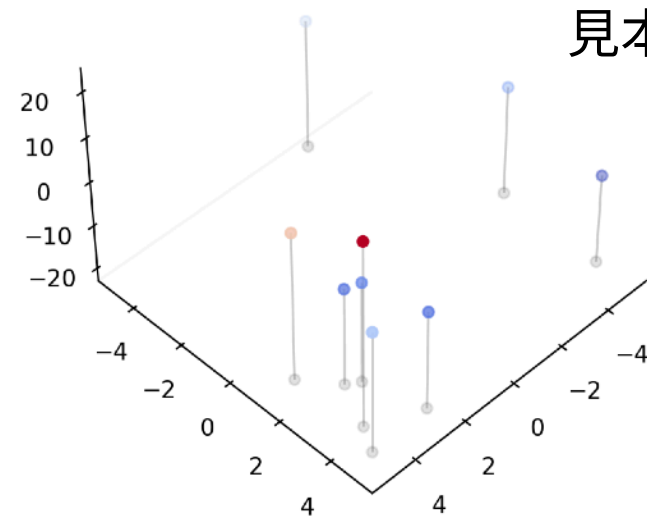
機械学習を使うのにどれくらいデータが必要ですか？

= 曲面の概形を知るのに見本点が何点必要ですか？

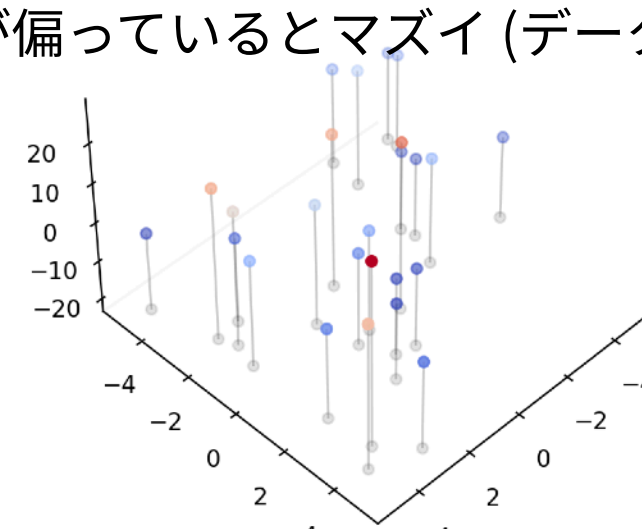


**曲面(入出力)が複雑だと
もっとデータ必要**

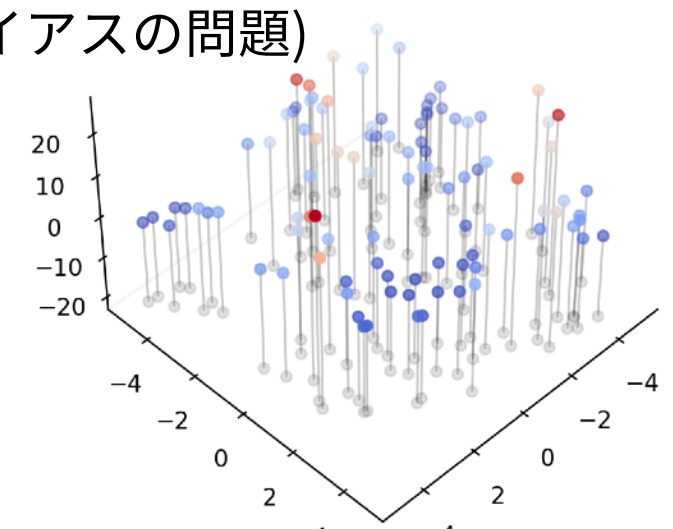
$N = 3^2 = 9$



$N = 5^2 = 25$

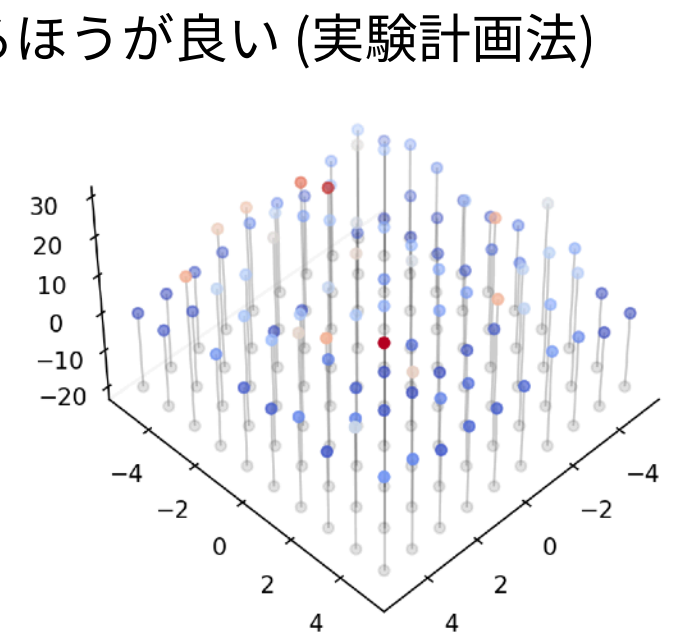
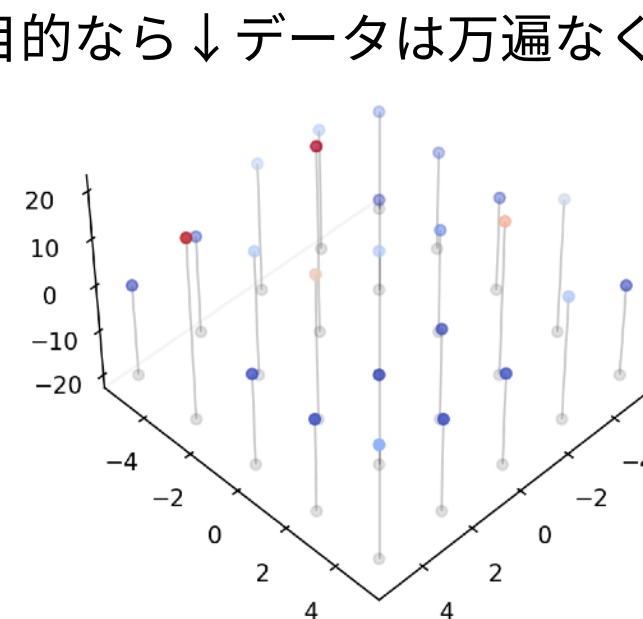
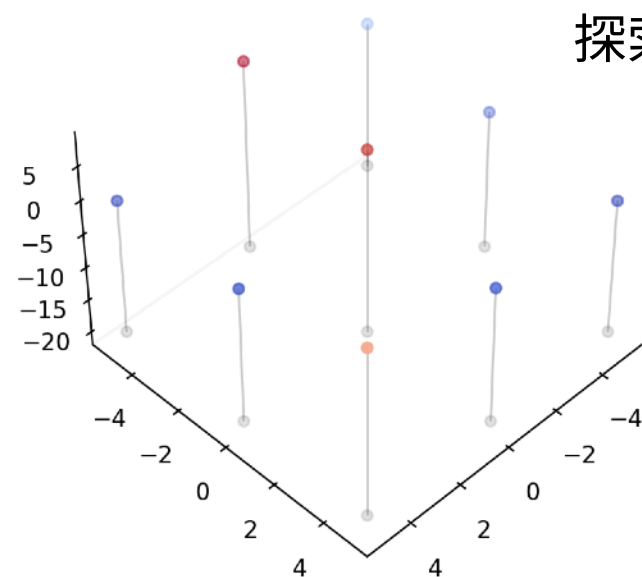


$N = 10^2 = 100$



見本例が偏っているとマズイ (データバイアスの問題)

探索が目的なら ↓ データは万遍なく取るほうが良い (実験計画法)

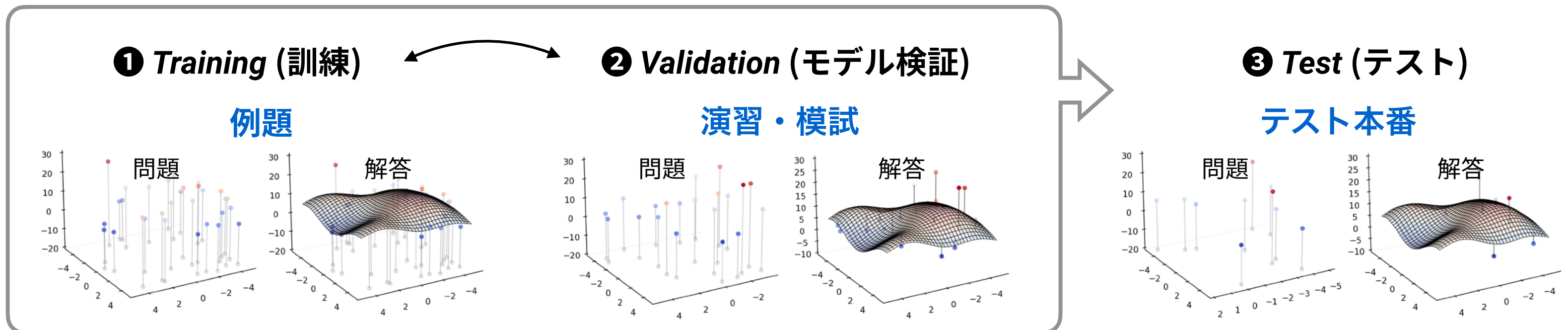


機械学習の予測が当たっているかの確認には落とし穴多数！

機械学習が「与えた見本例」を正しく予測できるのは**当たり前**

真の関心：「**見本例ではないデータ**」に対して正しく予測できるか？

- ✓ 予測精度も手元のデータから見積もるしかないので一般にこの判断は激ムズ
- ✓ データは手元にもうあるので意図しないカンニング事故(data leakage)がとても起きやすい…



半世紀前に生まれた「希望的な呼称」による幻想にご注意

現在の機械学習は一般に想像するSF的な「人工知能(AI)」とは**かなりかけ離れている**が、
「データを予測に変える」機能が**あまりに強力**なため、私たちの日常生活から今後の社会のカタチにまで影響を及ぼそうとしている…

「人工知能」「機械学習」などの**希望的な呼称**は本質をミスリードしやすいのでご注意を！

https://spectrum.ieee.org/files/11920/10_Spectrum_2021.pdf

SIGART Newsletter No. 57 April 1976

ARTIFICIAL INTELLIGENCE MEETS NATURAL STUPIDITY

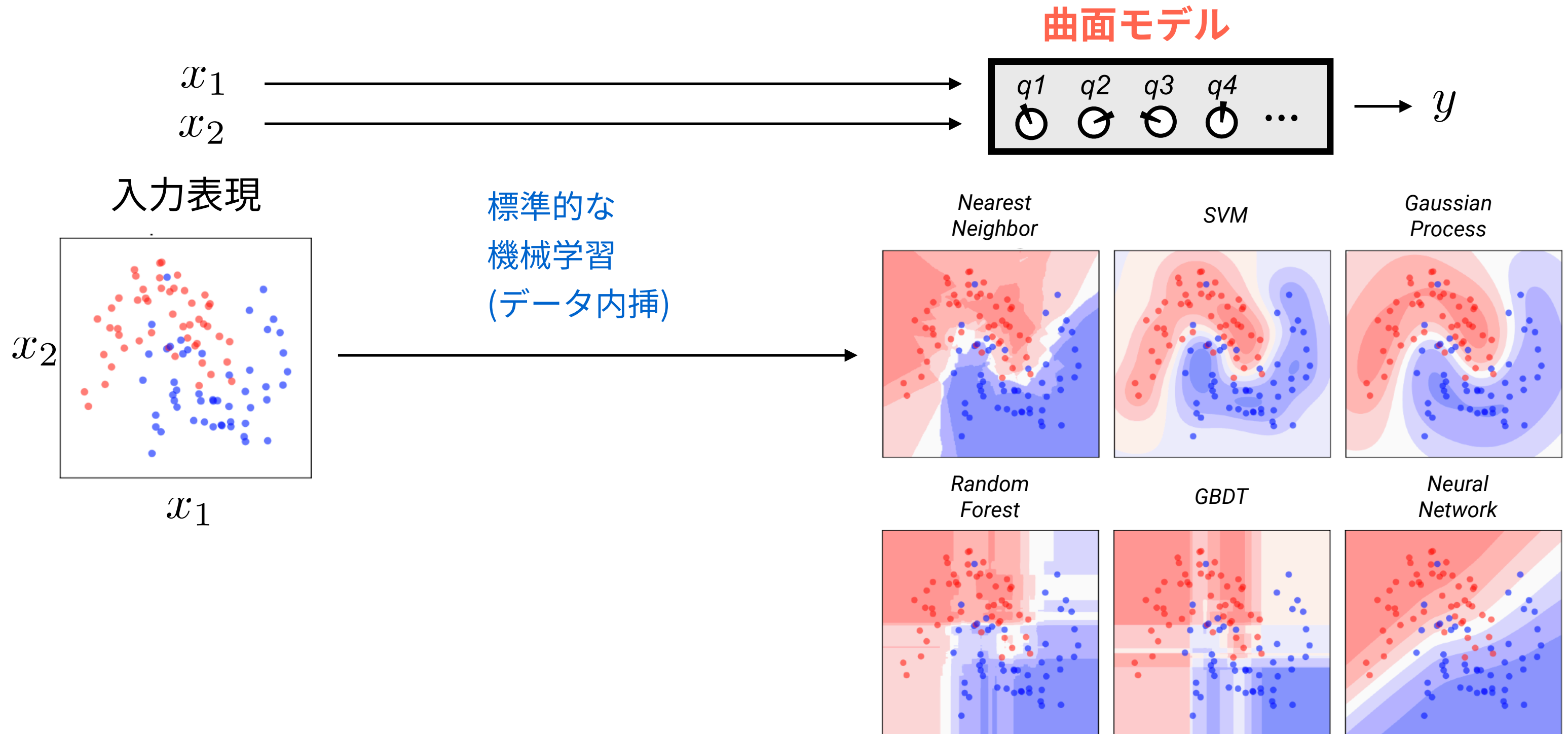
Drew McDermott

MIT AI Lab Cambridge, Mass 02139

As a field, artificial intelligence has always been on the border of respectability, and therefore on the border of crackpottery. Many critics <Dreyfus, 1972>, <Lighthill, 1973> have urged that we are over the border. We have been very defensive toward this charge, drawing ourselves up with dignity when it is made and folding the cloak of Science about us. On the other hand, in private, we have been justifiably proud of our willingness to explore weird ideas, because pursuing them is the only way to make progress.

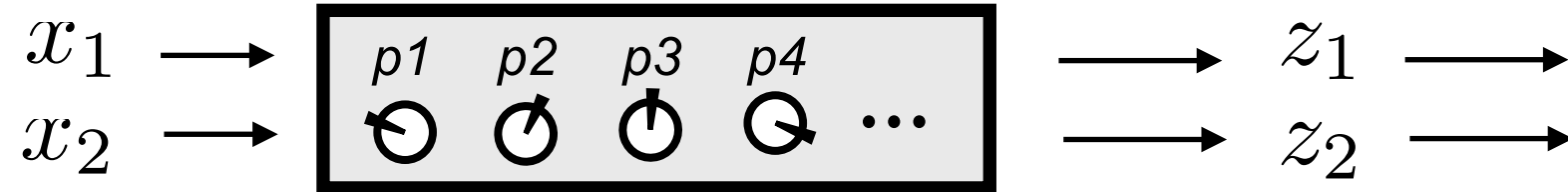


深層学習と表現学習

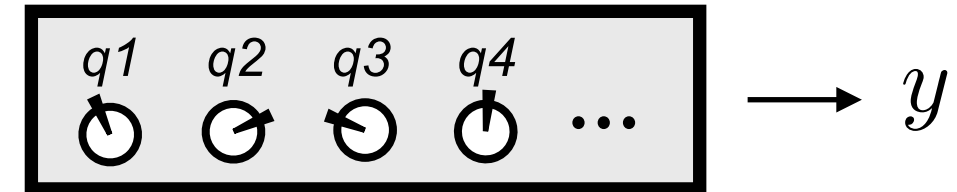


深層学習と表現学習

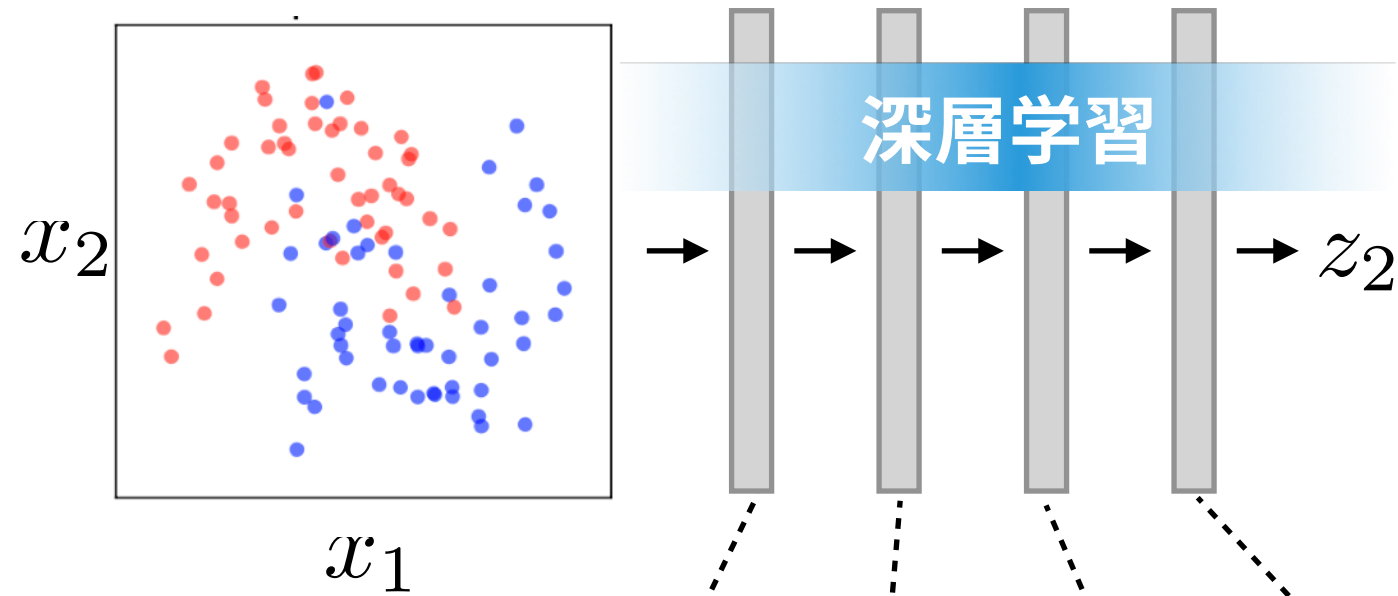
変数変換 (表現学習)



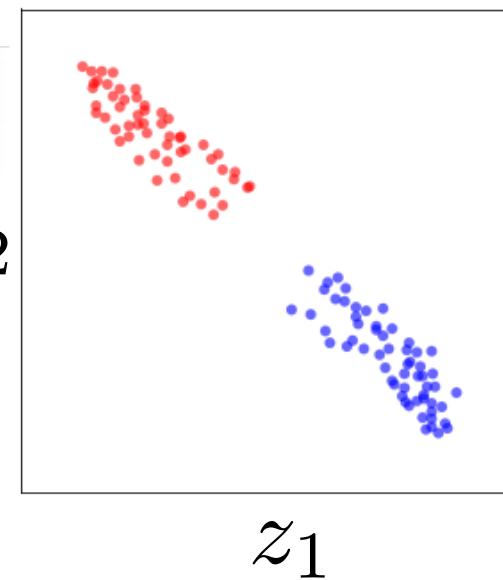
曲面モデル



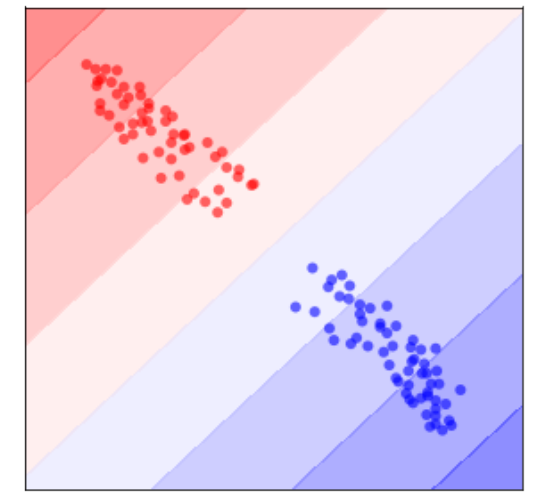
入力表現



良い表現

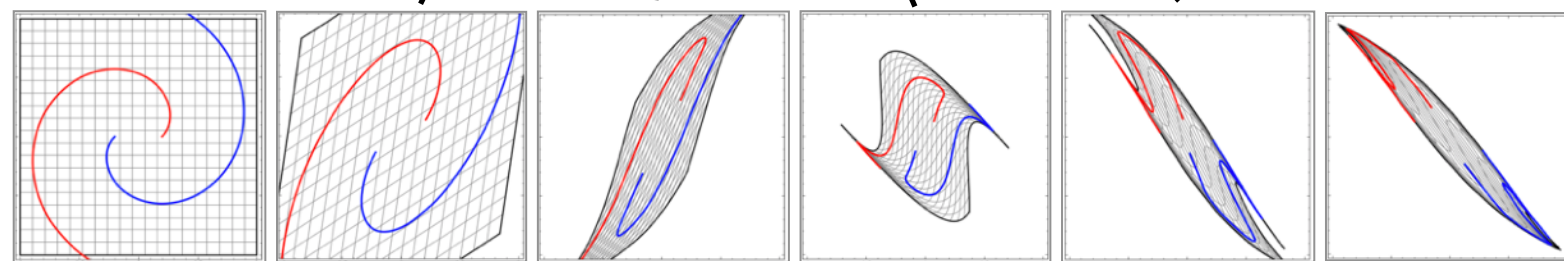


標準的な
機械学習
(データ内挿)



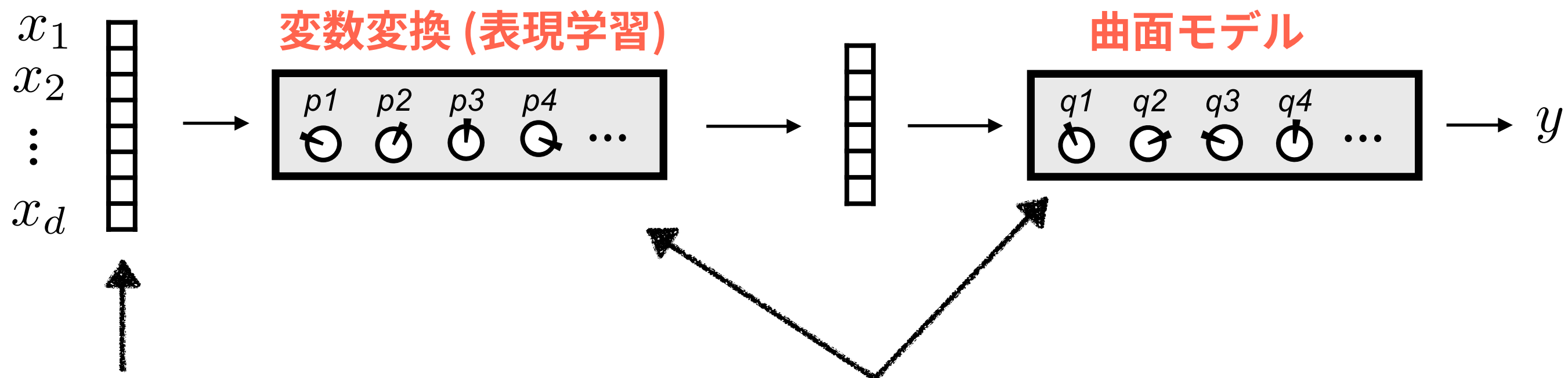
シンプル(線形)で十分

深層学習



データ内挿は入力表現ではなく「良い表現」で行う

困難① 現代の機械学習モデルは多量のデータを必要とする



① 高次元性：入力変数が多すぎ！

✓ 機械学習は入力されていない情報を全く考慮してくれない… (擬似相関リスク)

✓ とりあえず色々な変数を入れがち

画像そのままを入力する場合

20×20 ピクセルのカラー画像 → 1200変数

1000×1000 ピクセルのカラー画像 → 300万変数

② 過剰パラメタ化：パラメタ数が多すぎ！

画像 ResNet50: 2600万パラメタ
ResNet101: 4500万パラメタ
EfficientNet-B7: 6600万パラメタ
VGG19: 1億4400万パラメタ

言語 12-layer, 12-heads BERT: 1億1000万パラメタ
24-layer, 16-heads BERT: 3億3600万パラメタ
GPT-2 XL: 15億5800万パラメタ
GPT-3: 1750億パラメタ

困難② 羅生門効果とUnderspecification

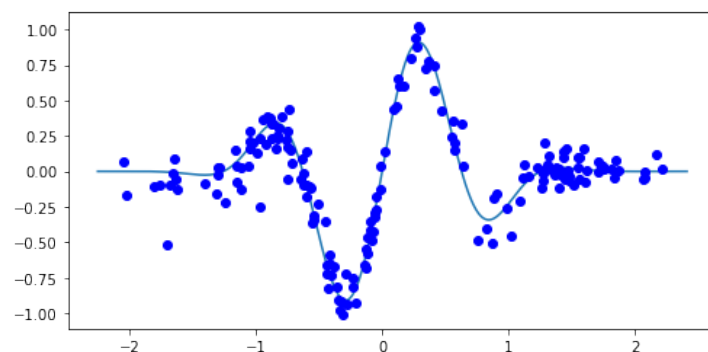
羅生門効果：良い機械学習モデルの多重性 (非一意性)

高い予測精度を持つ機械学習モデルは一つのデータセットからたくさん作れる！

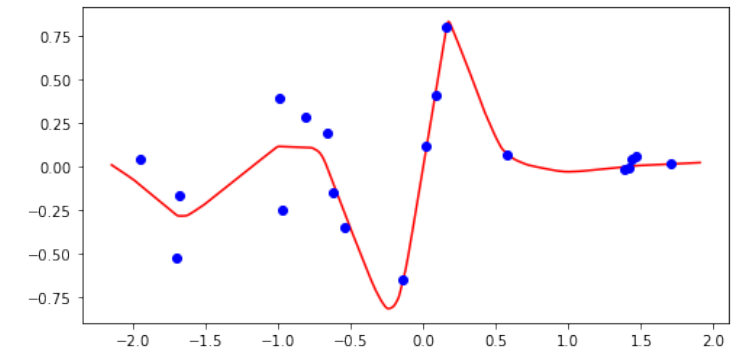
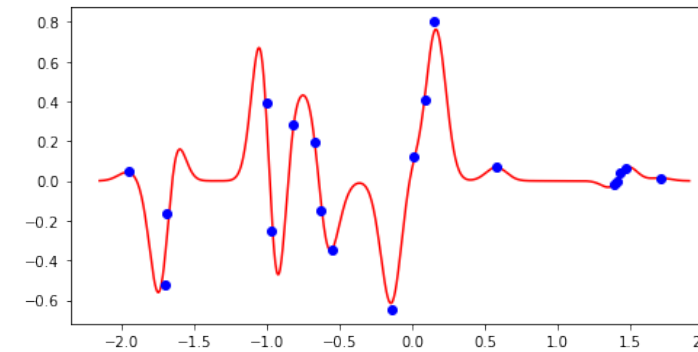
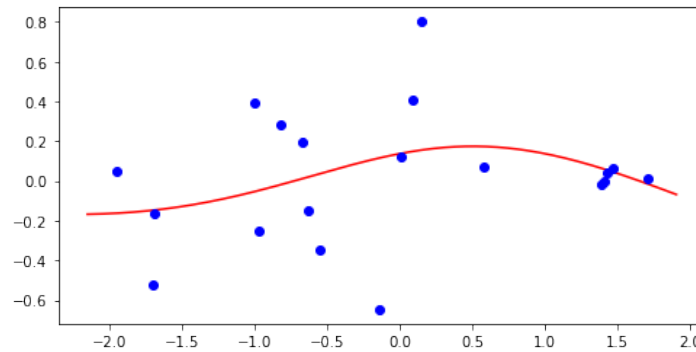
有限データから見積もる予測精度では指数的広さのモデル探索空間をspecifyしきれない…

- ✓ 「どのモデルが求める真実なの？」と考えてしまうと、まさに真実は「藪の中」…
→ 複数の手法による多角的解釈が鉄則 "all models are wrong but some models are useful"
- ✓ 実際には本質的にデータが足りてない(Underspecification)ことで多重性はさらに悪化

だいたいの方法で類似



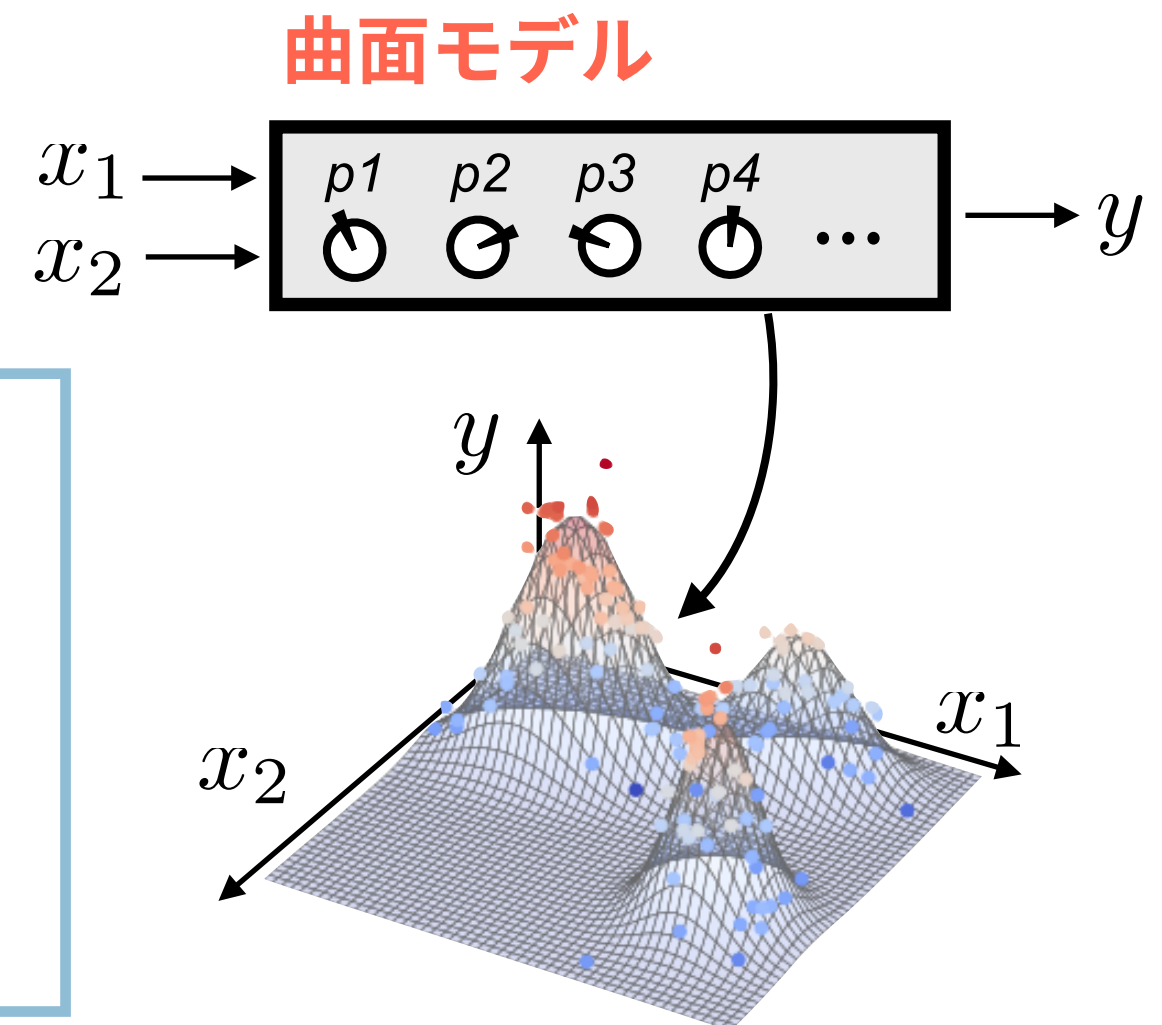
手法やモデルによって予測時の挙動にかなり差が出てしまう



現代の技術的関心はこの高次元性をどう手懐けるか

1. 確率的最適化・正則化 → モデルが大きい自由度の中で暴れまくらないよう動ける範囲を何とかして制御・制限・安定化する
2. 事前学習 (Warm Start)の転移 → 事前に得ておいたイイ感じのパラメタ初期値を使う
3. 帰納バイアスの設計

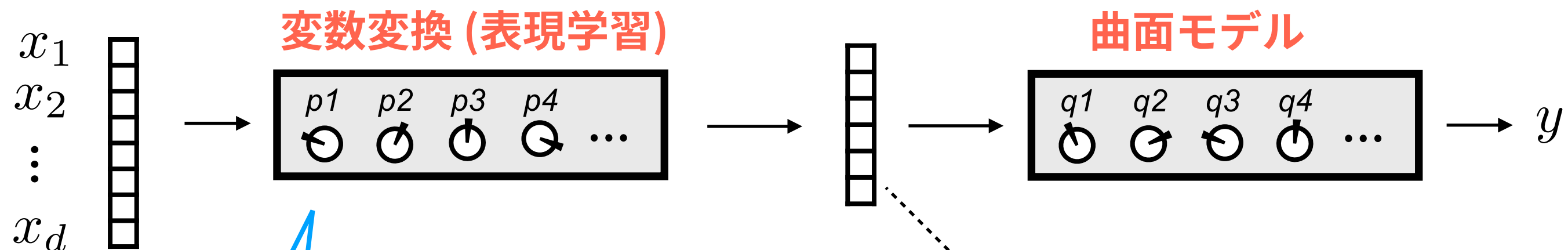
曲面モデルがどんな入出力関係でも表現できることが逆に擬似相関やUnderspecificationの問題を悪化させている



機械学習×化学：化学に適合した帰納バイアスのデザイン

化学的に妥当性を欠くようなモデルが**意図せず表現されてしまわないように**化学の知識や理論科学・計算化学の知見を総動員して**モデルの自由度を技術的に制限**する！

小サンプル問題の克服：大規模事前学習とその転移



収集可能な何らかの「大規模なデータ」に対して
本来のタスクとは別の「自己教師あり(SSL)」タスク
(Pretextタスク)を設計し変数変換部を事前に獲得する

SSL Pretextタスクの例 (人手の正解ラベルづけが不要)

- 文の単語をランダムに隠してそれを当てる
- 絵を90°/180°/270°にランダム回転させ角度を当てる
- 分子構造の一部を隠してそれを当てる

理想

一旦、変数変換が獲得できれば
個別事例では小サンプルでOK

「良い表現」が得られれば曲面モデルは
シンプル(線形)で良く **小サンプルでOK**

現実

結局、大規模データの取得が困難…

画像・音声・テキストなど華々しい成功例
のケースと違い、自然現象を扱う分野では
大規模データの取得が現状難しい…

機械学習×化学の真の問題

「予測ができる」ことは「理解」や「発見」ができることを直接は意味しない！！

下記はどれも機械学習でかなり高精度な予測ができますが、それは私たちがその仕組みを理解できたことを少しでも意味するのでしょうか？



どうやって見えているものを認識？



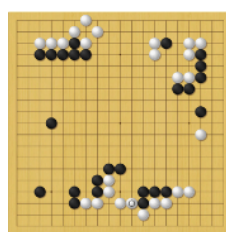
どうやって音波から言葉を理解？

“ありがとう”

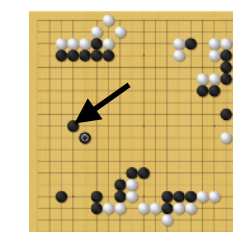
J'aime la
musique

どうやって多言語を理解？

I love music



どうやったら囲碁勝てる？

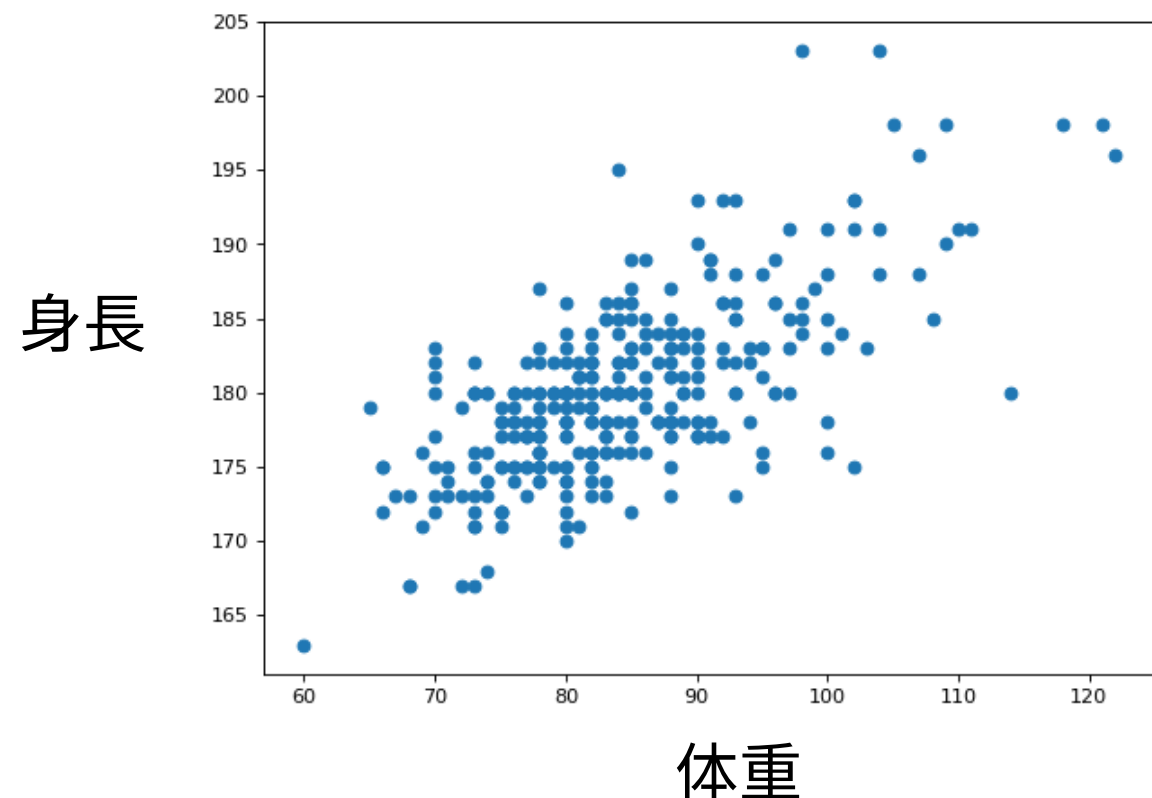


因果の理解には実験研究(介入研究)が必要不可欠

機械学習はあくまで**データの中の多次元相関を捉え、それによって予測する技術**
➡ 観察された相関が本当に**因果性**を含むのかを確かめるためには実験するしかない！

日本プロ野球開幕一軍選手の身長・体重データ

(2016年球団公式サイト選手データより自作)



「**体重を増やせば身長も伸びる**」が正しいかは
この観察データ**だけ**からは決して分からない

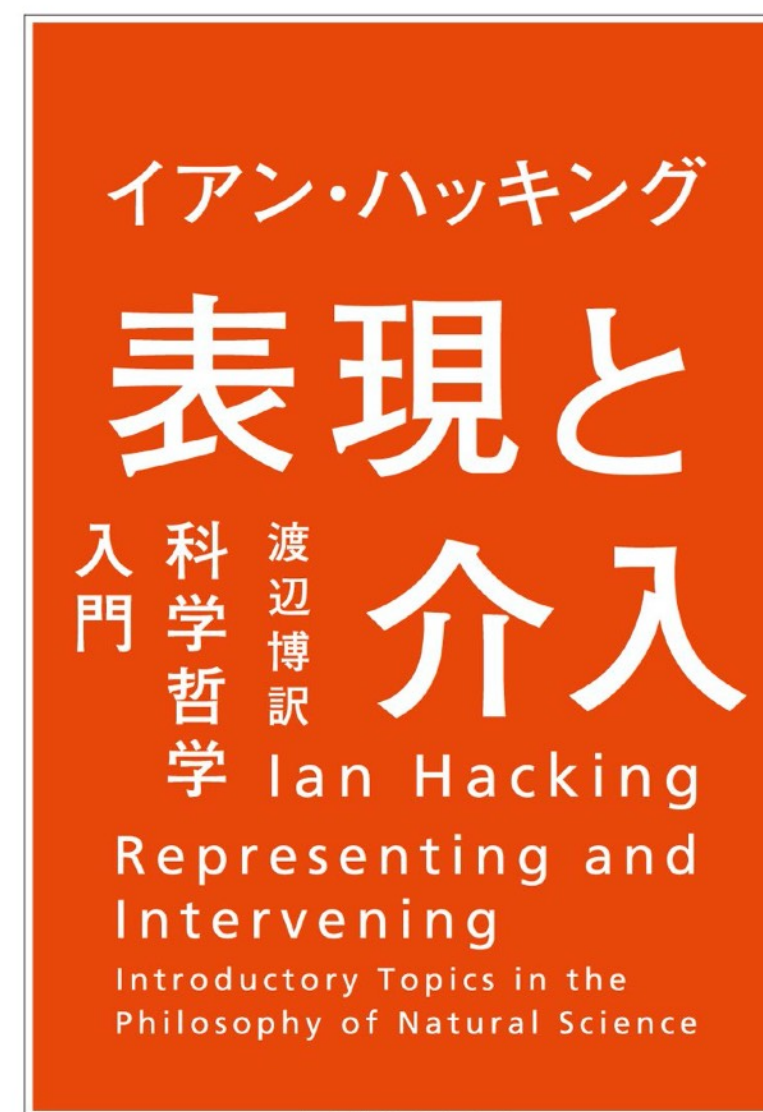
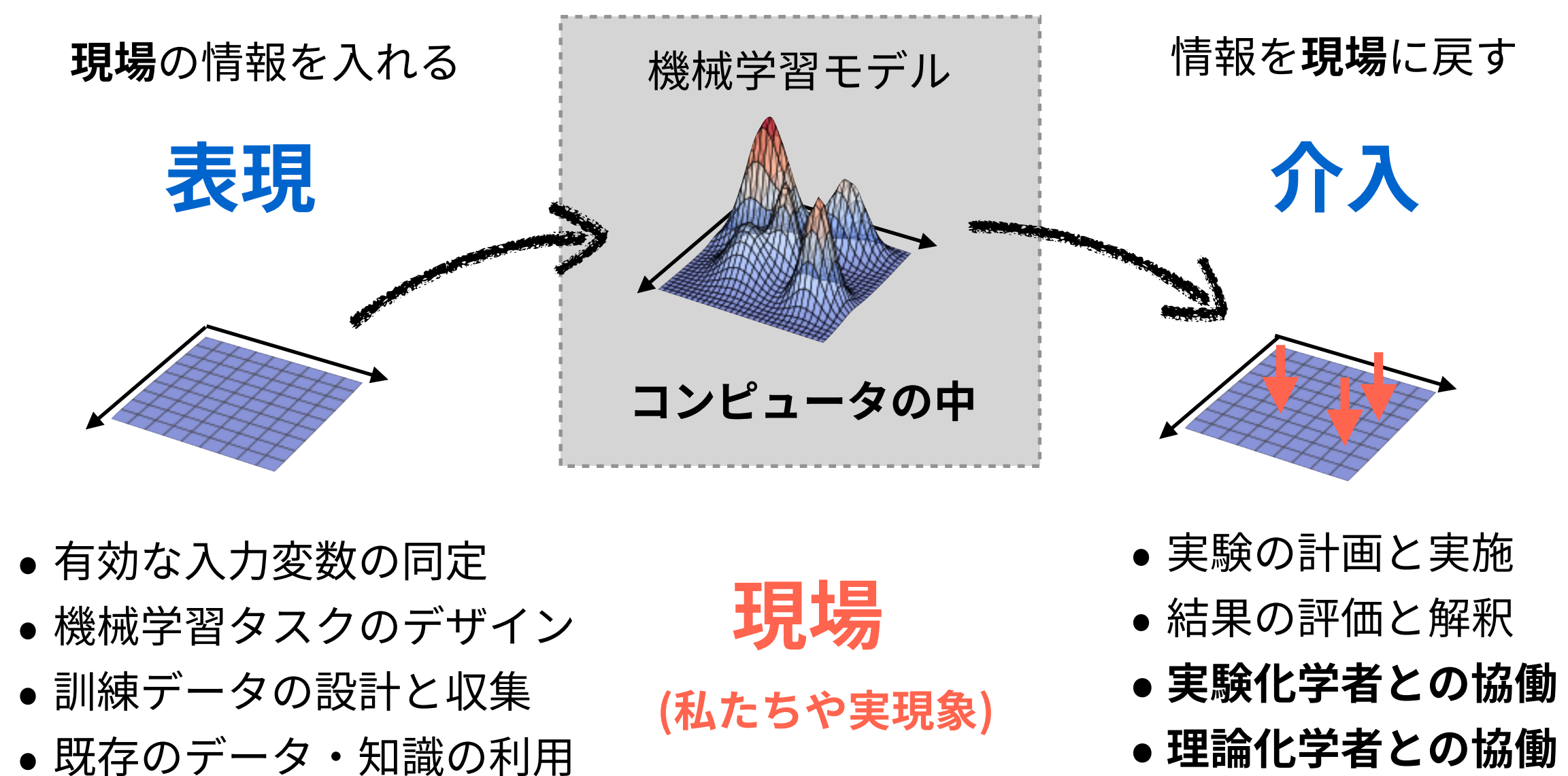
応用統計学の基本のキ

相関関係は必ずしも因果関係を意味しない

➡ 「**予測ができる**」ことは「**理解**」や「**発見**」が
できることを直接は意味しない！！

表現と介入：予測から理解・発見へ

事件はコンピュータ(機械学習)の中で起きてるんじゃない、**現場**で起きているんだ！ by 俺



教訓：科学研究とは結局人間の営み！

「理解」や「発見」したいのは機械ではなく私たち人間

つまり、自然法則の問題ではなく**私たち自身の精神と世界のあり方の問題**を問うことになる！

- **解釈性**：私たちの**ショボい認知能力**に収まるような「平易な理解」が求められている。
- **時間性**：**有限の時間**しか生きられない私たちに「発見」という体験をお膳立てするためのヒント出しが求められている。（人類絶滅のタイムリミット内に）
- **情報の部分性**：データにできる情報は**いつでも世界の情報量のほんのひとかけら**だけ。ゆく河の流れは絶えずして、しかももとの水にあらず。すべてを観測することはできない。
- **選択バイアス**：人間が一生懸命集めたデータはどうしたって**何らかの偏り**から逃れられない。与えられたデータの傾向を捉える機械学習の予測も同様にその偏りから逃れられない。
- **因果性の理解**：**「因果性」は直接観測できない**。人間がアクセスできるのは「相関」だけ！

まとめ：機械学習を自然現象の理解・発見に活用するとは？

必要な情報のうち、いつも偏った「一部」しかデータにはできない前提で、私たち自身の許容限界に見合う情報や示唆を得るために「データを予測に変える道具」をどう使えるか

- 新たな枠組み「**見本例によるプログラミング**」を「どこにどう使うか」のセンスが問われる。
「明示的な関係はよく分からないが入出力見本データは取れる」部分問題を熟考すること。
- 現状では「**本質的にはデータが足りてない**」場合がほとんどであり、専門家と機械学習屋が協働で分野の今までの知識や知見を生かし上手に「**帰納バイアス**」を設計する必要がある。
- 「因果性」は直接観測できないので、「**実際に実験によって確かめてみる**」介入が不可欠。
この検証ステップをどのようにデザイン・実現するかが非常に大切。
- 「**大規模データが得られる設定では非常に強力な技術**」なので近視眼的に今のところ手に入るデータだけで何とか場当たりの的に頑張り続ける以上の中長期的なデータ獲得戦略が大事。